Lecture Notes in Statistics

Edited by J. Berger, S. Fienberg, J. Gani,
K. Krickeberg, and B. Singer

48

G. Larry Bretthorst

Bayesian Spectrum Analysis and Parameter Estimation

Springer-Verlag

Permission To Distribute Electronically

On Wed Feb. 12, 1997 I wrote John Kimmel (Senior Editor, Statistics, Springer-Verlag):

To: jkimmel@worldnet.att.net Wed Feb 12 15:46:38 1997

Dear Mr. Kimmel,

About 9 years ago I wrote and published a book in your Lecture Notes series, Vol. 48, Bayesian Spectrum Analysis and Parameter Estimation. It has come to my attention that this book is now out of print. I still receive requests for copies of this book (although, not large numbers of them). I maintain an FTP/WWW site for distribution of materials on Bayesian Probability theory, and I was wondering if you, Springer, would mind if I posted a copy of my book. As Springer owns the copyrights to this book, I will not post it without permission. So my question is really two fold, does Springer plan to bring out a second printing and, if not, may I post a copy of it on the network?

Sincerely,

Larry Bretthorst, Ph.D.
Phone 314-362-9994

Later that day John Kimmel replayed:

Dear Dr. Bretthorst:

Lecture note volumes are rarely reprinted. Given that yours was published in 1988, I do not think that there would be enough volume to justify a reprint. You have our permission to make an electronic version available.

Your book seems to have been very popular. Would you be interested in a second edition or a more extensive monograph?
Best Regards,

John Kimmel

Springer-Verlag
25742 Wood Brook Rd.
Laguna Hills, CA 92653
U.S.A.

Phone: 714-582-6286
FAX: 714-348-0658
E-mail: jkimmel@worldnet.att.net

Author

G. Larry Bretthorst
Department of Chemistry, Campus Box 1134, Washington University
1 Brookings Drive, St. Louis, MO 63130, USA

Mathematics Subject Classification: 62F 15, 62Hxx

ISBN 0-387-96871-7 Springer-Verlag New York Berlin Heidelberg
ISBN 3-540-96871-7 Springer-Verlag Berlin Heidelberg New York

Printing and binding: Druckhaus Beltz, Hemsbach/Bergstr.
2847/3140-543210

To E. T. Jaynes

Preface

This work is essentially an extensive revision of my Ph.D. dissertation, [1]. It is primarily a research document on the application of probability theory to the parameter estimation problem. The people who will be interested in this material are physicists, economists, and engineers who have to deal with data on a daily basis; consequently, we have included a great deal of introductory and tutorial material. Any person with the equivalent of the mathematics background required for the graduate-level study of physics should be able to follow the material contained in this book, though not without effort.

From the time the dissertation was written until now (approximately one year) our understanding of the parameter estimation problem has changed extensively. We have tried to incorporate what we have learned into this book.

I am indebted to a number of people who have aided me in preparing this document: Dr. C. Ray Smith, Steve Finney, Juana Sanchez, Matthew Self, and Dr. Pat Gibbons who acted as readers and editors. In addition, I must extend my deepest thanks to Dr. Joseph Ackerman for his support during the time this manuscript was being prepared.

Last, I am especially indebted to Professor E. T. Jaynes for his assistance and guidance. Indeed it is my opinion that Dr. Jaynes should be a coauthor on this work, but when asked about this, his response has always been "Everybody knows that Ph.D. students have advisors." While his statement is true, it is essentially irrelevant; the amount of time and effort he has expended providing background material, interpretations, editing, and in places, writing this material cannot be overstated, and he deserves more credit for his effort than an "Acknowledgment."

St. Louis, Missouri, 1988 G. Larry Bretthorst

Contents

List of Figures

Chapter 1

INTRODUCTION

Experiments are performed in three general steps: first, the experiment must be designed; second, the data must be gathered; and third, the data must be analyzed. These three steps are highly idealized, and no clear boundary exists between them. The problem of analyzing the data is one that should be faced early in the design phase. Gathering the data in such a way as to learn the most about a phenomenon is what doing an experiment is all about. It will do an experimenter little good to obtain a set of data that does not bear directly on the model, or hypotheses, to be tested.

In many experiments it is essential that one does the best possible job in analyzing the data. This could be true because no more data can be obtained, or one is trying to discover a very small effect. Furthermore, thanks to modern computers, sophisticated data analysis is far less costly than data acquisition, so there is no excuse for not doing the best job of analysis that one can.

The theory of optimum data analysis, which takes into account not only the raw data but also the prior knowledge that one has to supplement the data, has been in existence – at least, as a well-formulated program – since the time of Laplace. But the resulting Bayesian probability theory (i.e., the direct application of probability theory as a method of inference) using realistic models has been little applied to spectral estimation problems and in science in general. Consequently, even though probability theory is well understood, its application and the orders of magnitude improvement in parameter estimates that its application can bring, are not. We hope to show the advantage of using probability theory in this way by developing a little of it and applying the results to some real data from physics and economics.

The basic model we are considering is always: we have recorded a discrete data

set $D = \{d_1, \cdots, d_N\}$, sampled from $y(t)$ at discrete times $\{t_1, \cdots, t_N\}$, with a model equation

$$d_i = y(t_i) = f(t_i) + e_i, \qquad (1 \leq i \leq N)$$

where $f(t_i)$ is the signal and e_i represents noise in the problem. *Different models correspond to different choices of the signal* $f(t)$. The most general model we will analyze will be of the form

$$f(t) = \sum_{j=1}^{m} B_j G_j(t, \{\omega\}).$$

The model functions, $G_i(t, \{\omega\})$, are functions of other parameters $\{\omega_1, \cdots, \omega_r\}$ which we label collectively $\{\omega\}$ (these parameters might be frequencies, chirp rates, decay rates, the time of some event, or any other quantities one could encounter).

We have not assumed the time intervals to be uniform, nor have we assumed the data to be drawn from some stationary Gaussian process. Indeed, in the most general formulation of the problem such considerations will be completely irrelevant. In the traditional way of thinking about this problem, one imagines that the data are one sample drawn from an infinite population of possible samples. One then uses probability only for the distribution of possible samples that could have been drawn – but were not. Instead, what we will do is to concentrate our attention on the actual data obtained, and use probability to make the "best" estimate of the parameters; i.e. the values that were realized when the data were taken.

We will concentrate on the $\{\omega\}$ parameters, and often consider the amplitudes $\{B\}$ as nuisance parameters. The basic question we would like to answer is: "What are the best estimates of the $\{\omega\}$ parameters one can make, independent of the amplitudes $\{B\}$ and independent of the noise variance?" We will solve this problem for the case where we have little prior information about the amplitudes $\{B\}$, the $\{\omega\}$ parameters, and the noise. Because we incorporate little prior information into the problem beyond the form of the model functions, the estimates of the amplitudes $\{B\}$ and the nonlinear $\{\omega\}$ parameters cannot differ greatly from the estimates one would obtain from least squares or maximum likelihood. However, using least squares or maximum likelihood would require us to estimate all parameters, interesting and non-interesting, simultaneously; thus one would have the computational problem of finding a global maximum in a space of high dimensionality.

By direct application of probability theory we will be able to remove the uninteresting parameters and see what the data have to tell us about the interesting ones, reducing the problem to one of low dimensionality, equal to the number of interesting

parameters. In a typical "small" problem this might reduce the search dimensions from ten to two; in one "large" problem the reduction was from thousands to six or seven. This represents many orders of magnitude reduction in computation, the difference between what is feasible, and what is not.

Additionally, the direct application of probability theory also tells us the accuracy of our estimates, which direct least squares does not give at all, and which maximum likelihood gives us only by a different calculation (sampling distribution of the estimator) which can be more difficult than the high-dimensional search one – and even then refers only to an imaginary class of different data sets, not the specific one at hand.

In Chapter 2, we analyze a time series which contains a single stationary harmonic signal plus noise, because it contains most of the points of principle that must be faced in the more general problem. In particular we derive the probability that a signal of frequency ω is present, regardless of its amplitude, phase, and the variance of the noise. We then demonstrate that the estimates one obtains using probability theory are a full order of magnitude better than what one would obtain using the discrete Fourier transform as a frequency estimator. This is not magic; we are able to understand intuitively why it is true, and also to show that probability theory has built-in automatic safety devices that prevent it from giving overoptimistic accuracy claims. In addition, an example is given of numerical analysis of real data illustrating the calculation.

In Chapter 3, we discuss the types of model equations used, introduce the concept of an orthonormal model, and derive a transformation which will take any nonorthonormal model into an orthonormal one. Using these orthonormal models, we then remove the simplifying assumptions that were made in Chapter 2, generalize the analysis to arbitrary model equations, and discuss a number of surprising features to illustrate the power and generality of the method, including an intuitive picture of model fitting that allows one to understand which parameters probability theory will estimate and why, in simple terms.

In Chapter 4 we calculate a number of posterior expectation values including the first and second moments, define a power spectral density, and we devise a procedure for estimating the nonlinear $\{\omega\}$ parameters.

In Chapter 5 we turn our attention to the problem of selecting the "best" model of a process. Although this problem sounds very different from the parameter estimation problem, it is essentially the same calculation. Here, we compute the relative posterior

probability of a model: this allows one to select the most probable model based on how well its parameters are estimated, and how well it fits the data.

In Chapter 6, we specialize the discussion to spectral estimates and, proceeding through stages, investigate the one-stationary-frequency problem and explicitly calculate the posterior probability of a simple harmonic frequency independent of its amplitude, phase and the variance of the noise, without the simplifying assumptions made in Chapter 2.

At that point we pause briefly to examine some of the assumptions made in the calculation and show that when these assumptions are violated by the data, the answers one obtains are still correct in a well-defined sense, but more conservative in the sense that the accuracy estimates are wider. We also compare uniform and nonuniform time sampling and demonstrate that for the single-frequency estimation problem, the use of nonuniform sampling intervals does not affect the ability to estimate a frequency. However, for apparently randomly sampled time series, aliases effectively do not exist.

We then proceed to solve the one-frequency-with-Lorentzian-decay problem and discuss a number of surprising implications for how decaying signals should be sampled. Next we examine the two stationary frequency problem in some detail, and demonstrate that (1) the ability to estimate two close frequencies is essentially independent of the separation as long as that separation is at least one Nyquist step $|\omega_1 - \omega_2| \geq 2\pi/N$; and (2) that these frequencies are still resolvable at separations corresponding to less than one half step, where the discrete Fourier transform shows only a single peak.

After the two-frequency problem we discuss briefly the multiple nonstationary frequency estimation problem. In Chapter 3 Eq. (3.17) we derive the joint posterior probability of multiple stationary or nonstationary frequencies independent of their amplitude and phase and independent of the noise variance. Here we investigate some of the implications of these formulas and discuss the techniques and procedures needed to apply them effectively.

In Chapter 7, we apply the theory to a number of real time series, including Wolf's relative sunspot numbers, some NMR (nuclear magnetic resonance) data containing multiple close frequencies with decay, and to economic time series which have large trends. The most spectacular results obtained to date are with NMR data, because here prior information tells us very accurately what the "true" model must be.

Equally important, particularly in economics, is the way probability theory deals with trend. Instead of seeking to eliminate the trend from the data (which is known to

introduce spurious artifacts that distort the information in the data), we seek instead to eliminate the effect of trend from the final conclusions, leaving the data intact. This proves to be not only a safer, but also a more powerful procedure than detrending the data. Indeed, it is now clear that many published economic time series have been rendered nearly useless because the data have been detrended or seasonally adjusted in an irreversible way that destroys information which probability theory could have extracted from the raw, unmutilated data.

In the last example we investigate the use of multiple measurements and show that probability theory can continue to obtain the standard $\sqrt{n}$ improvement in parameter estimates under much wider conditions than averaging. The analyses presented in Chapter 7 will give the reader a better feel for the types of applications and complex phenomena which can be investigated easily using Bayesian techniques.

1.1 Historical Perspective

Comprehensive histories of the spectral analysis problem have been given recently by Robinson [2] and Marple [3]. We sketch here only the part of it that is directly ancestral to the new work reported here. The problem of determining a frequency in time sampled data is very old; the first astronomers were trying to solve this problem when they attempted to determine the length of a year or the period of the moon. Their methods were crude and consisted of little more than trying to locate the maxima or the nodes of an approximately periodic function. The first significant advance in the frequency estimation problem occurred in the early nineteenth century, when two separate methods of analyzing the problem came into being: the use of probability theory, and the use of the Fourier transform.

Probabilistic methods of dealing with the problem were formulated in some generality by Laplace [4] in the late 18th century, and then applied by Legendre and Gauss [5] [6] who first used (or at least first published) the method of least squares to estimate model parameters in noisy data. In this procedure some idealized model signal is postulated and the criterion of minimizing the sum of the squares of the "residuals" (the discrepancies between the model and the data) is used to estimate the model parameters. In the problem of determining a frequency, the model might be a single cosine with an amplitude, phase, and frequency, contaminated by noise with an unknown variance. Generally one is not interested in the amplitude, phase,

or noise variance; ideally one would like to formulate the problem in such a way that only the frequency remains, but this is not possible with direct least squares, which requires us to fit all the model parameters. The method of least squares may be difficult to use in practice; in principle it is well understood. In the case of Gaussian noise, the least squares estimates are simply the parameter values that maximize the probability that we would obtain the data, if a model signal was present with those parameters.

The spectral method of dealing with this problem also has its origin in the early part of the nineteenth century. The Fourier transform is one of the most powerful tools in analysis, and its discrete analogue is by definition the spectrum of the time sampled data. How this is related to the spectrum of the original time series is, however, a nontrivial technical problem whose answer is different in different circumstances. Using the discrete Fourier transform of the data as an estimate of the "true" spectrum is, intuitively, a natural thing to do: after all, the discrete Fourier transform is the spectrum of the noisy time sampled series, and when the noise goes away the discrete Fourier transform is the spectrum of the sampled "true" series, but calculating the spectrum of a series and estimating a frequency are very different problems. One of the things we will attempt to do is to do is to exhibit the exact conditions under which the discrete Fourier transform is an optimal frequency estimator.

With the introduction (or rather, rediscovery [7], [8], [9]) of the fast Fourier transform by Cooley and Tukey [10] in 1965 and the development of computers, the use of the discrete Fourier transform as a frequency and power spectral estimator has become very commonplace. Like the method of least squares, the use of discrete Fourier transform as a frequency estimator is well understood. If the data consist of a signal plus noise, then by linearity the Fourier transform will be the signal transform plus a noise transform. If one has plenty of data the noise transform will be, usually, a function of frequency with slowly varying amplitude and rapidly varying phase. If the peak of the signal transform is larger than the noise transform, the added noise does not change the location of the peak very much. One can then estimate the frequency from the location of the peak of the data transform, as intuition suggests.

Unfortunately, this technique does not work well when the signal-to-noise ratio of the data is small; then we need probability theory. The technique also has problems when the signal is other than a simple harmonic frequency: then the signal has some type of structure [for example Lorentzian or Gaussian decay, or chirp: a chirped signal has the form $\cos(\theta + \omega t + \alpha t^2)$]. The peak will then be spread out relative to a simple

harmonic spectrum. This allows the noise to interfere with the parameter estimation problem much more severely, and probability theory becomes essential. Additionally, the Fourier transform is not well defined when the data are nonuniform in time, even though the problem of frequency estimation is not essentially changed.

Arthur Schuster [11] introduced the periodogram near the beginning of this century, merely as an intuitive *ad hoc* method of detecting a periodicity and estimating its frequency. The periodogram is essentially the squared magnitude of the discrete Fourier transform of the data $D \equiv \{d_1, d_2, \cdots, d_N\}$ and can be defined as

$$C(\omega) = \frac{1}{N}\Big[R(\omega)^2 + I(\omega)^2\Big] = \frac{1}{N}|\sum_{j=1}^{N} d_j \mathrm{e}^{i\omega t_j}|^2, \tag{1.1}$$

where $R(\omega)$, and $I(\omega)$ are the real and imaginary parts of the sum [Eqs. (2.4), and (2.5) below], and N is the total number of data points. The periodogram remains well defined when the frequency ω is allowed to vary continuously or when the data are nonuniform. This avoids one of the potential drawbacks of using this method but does not aid in the frequency estimation problem when the signal is not stationary. Although Schuster himself had very little success with it, more recent experience has shown that regardless of its drawbacks, indeed the discrete Fourier transform or the periodogram does yield useful frequency estimates under a wide variety of conditions. Like least squares, Fourier analysis alone does not give an indication of the accuracy of the estimates of spectral density, although the width of a sharp peak is suggestive of the accuracy of determination of the position of a very sharp line.

In the 160 years since the introduction of the spectral and probability theory methods no particular connection between them had been noted, yet each of these methods seems to function well in some conditions. That these methods could be very closely related (from some viewpoints essentially the same) was shown when Jaynes [12] derived the periodogram directly from the principles of probability theory and demonstrated it to be, a "sufficient statistic" for inferences about a single stationary frequency or "signal" in a time sampled data set, when a Gaussian probability distribution is assigned for the noise. That is, starting with the same probability distribution for the noise that had been used for maximum likelihood or least squares, the periodogram was shown to be the only function of the data needed to make estimates of the frequency; i.e. it summarizes all the information in the data that is relevant to the problem.

In this work we will continue the analysis started by Jaynes and show that when the noise variance σ^2 is known, the conditional posterior probability density of a

frequency ω given the data D, the noise variance σ^2, and the prior information I is simply related to the periodogram:

$$P(\omega|D,\sigma,I) \propto \exp\left\{\frac{C(\omega)}{\sigma^2}\right\}. \tag{1.2}$$

Thus, we will have demonstrated the relation between the two techniques. Because the periodogram, and therefore the Fourier transform, will have been derived from the principles of probability theory we will be able to see more clearly under what conditions the discrete Fourier transform of the data is a valid frequency estimator and the proper way to extract optimum estimates from it. Also, from (1.2) we will be able to assess the accuracy of our estimates, which neither least squares, Fourier analysis, nor maximum likelihood give directly.

The term "spectral analysis" has been used in the past to denote a wider class of problems than we shall consider here; often, one has taken the view that the entire time series is a "stochastic process" with an intrinsically continuous spectrum, which we seek to infer. This appears to have been the viewpoint underlying the work of Schuster, and of Blackman-Tukey noted in the following sections. For an account of the large volume of literature on this version of the spectral estimation problem, we refer the reader to Marple [3].

The present work is concerned with what Marple calls the "parameter estimation method". Recent experience has taught us that this is usually a more realistic way of looking at current applications; and that when the parameter estimation approach is based on a correct model it can achieve far better results than can a "stochastic" approach, because it incorporates cogent prior information into the calculation. In addition, the parameter estimation approach proves to be more flexible in ways that are important in applications, adapting itself easily to such complicating features as chirp, decay, or trend.

1.2 Method of Calculation

The basic reasoning used in this work will be a straightforward application of Bayes' theorem: denoting by $P(A|B)$ the conditional probability that proposition A is true, given that proposition B is true, Bayes' theorem is

$$P(H|D,I) = \frac{P(H|I)P(D|H,I)}{P(D|I)}. \tag{1.3}$$

It is nothing but the probabilistic statement of an almost trivial fact: Aristotelian logic is commutative. That is, the propositions

$$HD \quad = \quad \text{“ Both } H \text{ and } D \text{ are true”}$$

$$DH \quad = \quad \text{“Both } D \text{ and } H \text{ are true”}$$

say the same thing, so they must have the same truth value in logic and the same probability, whatever our information about them. In the product rule of probability theory, we may then interchange H and D

$$P(H,D|I) = P(D|I)P(H|D,I) = P(H|I)P(D|H,I)$$

which is Bayes' theorem. In our problems, H is any hypothesis to be tested, D is the data, and I is the prior information. In the terminology of the current statistical literature, $P(H|D,I)$ is called the posterior probability of the hypothesis, given the data and the prior information. This is what we would like to compute for several different hypotheses concerning what systematic “signal” is present in our data. Bayes' theorem tells us that to compute it we must have three terms: $P(H|I)$ is the prior probability of the hypothesis (given only our prior information), $P(D|I)$ is the prior probability of the data (this term will always be absorbed into a normalization constant and will not change the conclusions within the context of a given model, although it does affect the relative probabilities of different models) and $P(D|H,I)$ is called the direct probability of the data, given the hypothesis and the prior information. The direct probability is called the “sampling distribution” when the hypothesis is held constant and one considers different sets of data, and it is called the “likelihood function” when the data are held constant and one varies the hypothesis. Often, a prior probability distribution is called simply a “prior”.

In a specific Bayesian probability calculation, we need to “define our model”; i.e. to enumerate the set $\{H_1, H_2, \cdots\}$ of hypotheses concerning the systematic signal in the model, that is to be tested by the calculation. A serious weakness of all Fourier transform methods is that they do not consider this aspect of the problem. In the widely used Blackman-Tukey [13] method of spectrum analysis, for example, there is no mention of any model or any systematic signal at all. In the problems we are considering, specification of a definite model (i.e. stating just what prior information we have about the phenomenon being observed) is essential; the information we can extract from the data depends crucially on which model we analyze.

In our problems, therefore, the Blackman-Tukey method, which does not even have the concept of a signal, much less a signal-to-noise ratio, would be inappropriate. Bayesian analysis based on a good model can achieve orders of magnitude better sensitivity and resolution. Indeed, one of our main new results is the very great improvement in resolution that can be achieved by replacing an unrealistic model by a realistic one.

In the most general model we will analyze, the hypothesis H will be of the form

$$H \equiv \text{“}f(t) = \sum_{j=1}^{m} B_j G_j(t, \{\omega\})\text{”}$$

where $f(t)$ is some analytic representation of the time series, $G_j(t, \{\omega\})$ is one particular model function (for example a sinusoid or trend), B_j is the amplitude with which G_j enters the model, and $\{\omega\}$ is a set of frequencies, decay rates, chirp rates, trend rate, or any other parameters of interest.

In the problem we are considering we focus our attention on the $\{\omega\}$ parameters. Although we will calculate the expectation value of the amplitudes $\{B\}$ we will not generally be interested in them. We will seek to formulate the posterior probability density $P(\{\omega\}|D, I)$ independently of the amplitudes $\{B\}$.

The principles of probability theory uniquely determine how this is to be done. Suppose ω is a parameter of interest, and B is a “nuisance parameter” that we do not, at least at the moment, need to know. What we want is $P(\omega|D, I)$, the posterior probability (density) of ω. This may be calculated as follows: first calculate the joint posterior probability density of ω and B by Bayes' theorem:

$$P(\omega, B|D, I) = P(\omega, B|I)\frac{P(D|\omega, B, I)}{P(D|I)}$$

and then integrate out B, obtaining the marginal posterior probability density for ω:

$$P(\omega|D, I) = \int dB P(\omega, B|D, I)$$

which expresses what the data and prior information have to tell us about ω, regardless of the value of B.

Although integration over the nuisance parameters may look a little strange at first glance, it is easily demonstrated to be a straightforward application of the sum rule of probability theory: the probability that one of several mutually exclusive propositions is true, is the sum of their separate probabilities. Suppose for simplicity that B is a discrete variable taking on the values $\{B_1, \cdots, B_n\}$ and we know that when the data

were taken only one value of B was realized; but we do not know which value. We can compute $P(\omega, \sum_{i=1}^{n} B_i|D, I)$ where the symbol "+" or "$\sum$" inside a probability symbol means the Boolean "or" operation [read this as the probability of (ω and B_1) or (ω and B_2)$\cdots$]. Using the sum rule this probability may be written

$$\begin{aligned} P(\omega, B_1 + \sum_{i=2}^{n} B_i|D, I) &= P(\omega, B_1|D, I) \\ &+ P(\omega, \sum_{i=2}^{n} B_i|D, I)[1 - P(\omega, B_1|\sum_{i=2}^{n} B_i D, I)]. \end{aligned}$$

The last term $P(\omega, B_1|\sum_{i=2}^{n} B_i D, I)$ is zero: only one value of B could be realized. We have

$$P(\omega, B_1 + \sum_{i=2}^{n} B_i|D, I) = P(\omega, B_1|D, I) + P(\omega, \sum_{i=2}^{n} B_i|D, I)$$

and repeated application of the sum rule gives

$$P(\omega, \sum_{i=1}^{n} B_i|D, I) = \sum_{i=1}^{n} P(\omega, B_i|D, I).$$

When the values of B are continuous the sums go into integrals and one has

$$P(\omega|D, I) = \int dB P(\omega, B|D, I), \tag{1.4}$$

the given rule. The term on the left is called the marginal posterior probability density function of ω, and it takes into account all possible values of B regardless of which actual value was realized. We have dropped the reference to B specifically because this distribution no longer depends on one specific value of B; it depends rather on all of them.

We discuss these points further in Appendices A, B, and C where we show that this procedure is similar to, but superior to, the common practice of estimating the parameter B from the data and then constraining B to that estimate.

In the following chapter we consider the simplest nontrivial spectral estimation model

$$f(t) = B_1 \cos \omega t + B_2 \sin \omega t$$

and analyze it in some depth to show some elementary but important points of principle in the technique of using probability theory with nuisance parameters and "uninformative" priors.

Chapter 2

SINGLE STATIONARY SINUSOID PLUS NOISE

2.1 The Model

We begin the analysis by constructing the direct probability, $P(D|H,I)$. We think of this as the likelihood of the parameters, because it is its dependence on the model parameters which concerns us here. The time series $y(t)$ we are considering is postulated to contain a single stationary harmonic signal $f(t)$ plus noise $e(t)$. The basic model is always: we have recorded a discrete data set $D = \{d_1, \cdots, d_N\}$; sampled from $y(t)$ at discrete times $\{t_1, \cdots, t_N\}$; with a model equation

$$d_i = y(t_i) = f(t_i) + e_i, \qquad (1 \leq i \leq N).$$

As already noted, *different models correspond to different choices of the signal* $f(t)$. We repeat the analysis originally done by Jaynes [12] using a different, but equivalent, set of model functions. We repeat this analysis for three reasons: first, by using a different formulation of the problem we can see how to generalize to multiple frequencies and more complex models; second, to introduce a different prior probability for the amplitudes, which simplifies the calculation but has almost no effect on the final result; and third, to introduce and discuss the calculation techniques without the complex model functions confusing the issues.

The model for a simple harmonic frequency may be written

$$f(t) = B_1 \cos(\omega t) + B_2 \sin(\omega t) \tag{2.1}$$

which has three parameters (B_1, B_2, ω) that may be estimated from the data. The

model used by Jaynes [12] was the same, but expressed in polar coordinates:

$$f(t) = B\cos(\omega t + \theta)$$

$$B = \sqrt{B_1^2 + B_2^2}$$

$$\tan\theta = -\frac{B_2}{B_1}$$

$$dB_1 dB_2 d\omega = B dB d\theta d\omega.$$

It is the factor B in the volume elements which is treated differently in the two calculations. Jaynes used a prior probability that initially considered equal intervals of θ and B to be equally likely, while we shall use a prior that initially considers equal intervals of B_1 and B_2 to be equally likely.

Of course, neither choice fully expresses all the prior knowledge we are likely to have in a real problem. This means that the results we find are conservative, and in a case where we have quite specific prior information about the parameters, we would be able to do somewhat better than in the following calculation. However, the differences arising from different prior probabilities are small provided we have a reasonable amount of data. For a detailed discussion and derivation of the prior probabilities used in this chapter, see Appendix A. In addition, in Appendix D we show explicitly that the prior used by Jaynes is more conservative for frequency estimation than the uniform prior we use, but when the signal-to-noise ratio is large the effect of this uninformative prior is completely negligible.

2.2 The Likelihood Function

To construct the likelihood we take the difference between the model function, or "signal", and the data. If we knew the true signal, then this difference would be just the noise. Then if we knew the probability of the noise we could compute the direct probability or likelihood. We wish to assign a noise prior probability density which is consistent with the available information about the noise. The prior should be as uninformative as possible to prevent us from "seeing" things in the data which are not there.

To derive the prior probability for the noise is a problem that can be approached in various ways. Perhaps the most general one is to view it as a simple application of the principle of maximum entropy. Let $P(e|I)$ stand for the probability that the

noise has value "e" given the prior information I. Then, assuming the second moment of the noise (i.e. the noise power) is known, the entropy functional which must be maximized is given by

$$-\int_{-\infty}^{+\infty} P(e|I)\log P(e|I)de - \lambda\int_{-\infty}^{+\infty} e^2 P(e|I)de - \beta\int_{-\infty}^{+\infty} P(e|I)de$$

where λ is the Lagrange multiplier associated with the second moment, and β is the multiplier for normalization. The solution to this standard maximization problem is

$$P(e|\lambda, I) = (\lambda/\pi)^{\frac{1}{2}}\exp\left\{-\lambda e^2\right\}.$$

Adopting the notation $\lambda = (2\sigma^2)^{-1}$, where σ^2 is the second moment, assumed known, we have

$$P(e|\sigma, I) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{e^2}{2\sigma^2}\right\}.$$

This is a Gaussian distribution, and when σ is taken as the RMS noise level, it is the least informative prior probability density for the noise that is consistent with the given second moment. By least informative we mean that: if any of our assumptions had been different and we used that information in maximum entropy to derive a new prior probability for the noise, then for a given σ, that new probability density would be less spread out, thus our accuracy estimates would be narrowed. Thus, in the calculations below, we will be claiming less accuracy than would be justified had we included those additional effects in deriving the prior probability for the noise. The point is discussed further in Chapter 5. In Chapter 6 we demonstrate (numerically) the effects of violating the assumptions that will go into the calculation. All of these "conservative" features are safety devices which make it impossible for the theory to mislead us by giving overoptimistic results.

Having the prior probability for the noise, and adopting the notation: e_i is the noise at time t_i, we apply the product rule of probability theory to obtain the probability that we would obtain a set of noise values $\{e_1, \cdots, e_N\}$: supposing the e_i independent in the sense that $P(e_i|e_j, \sigma, I) = P(e_i|\sigma, I)$ this is given by

$$P(e_1, \cdots, e_N|\sigma, I) \propto \prod_{i=1}^{N}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{e_i^2}{2\sigma^2})\right],$$

in which the independence of different e_i is also a safety device to maintain the conservative aspect. But if we have definite prior evidence of dependence, i.e. correlations, it is a simple computational detail to take it into account as noted later.

Other rationales for this choice exist in other situations. For example, if the noise is known to be the result of many small independent effects, the central limit theorem of probability theory leads to the Gaussian form independently of the fine details, even if the second moment is not known. For a detailed discussion of why and when a Gaussian distribution should be used for the noise probability, see the original paper by Jaynes [12]. Additionally, the book of Jaynes' collected papers contains a discussion of the principle of maximum entropy and much more [14].

If we have the true model, the difference between the data d_i and the model f_i is just the noise. Then the direct probability that we should obtain the data $D = \{d_1 \cdots d_N\}$, given the parameters, is proportional to the likelihood function:

$$P(D|B_1, B_2, \omega, \sigma, I) \propto L(B_1, B_2, \omega, \sigma) = \prod_{i=1}^{N} \sigma^{-1} \exp\{-\frac{1}{2\sigma^2}[d_i - f(t_i)]^2\}$$

$$L(B_1, B_2, \omega, \sigma) = \sigma^{-N} \times \exp\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} [d_i - f(t_i)]^2\}. \tag{2.2}$$

The usual way to proceed is to fit the sum in the exponent. Finding the parameter values which minimize this sum is called "least squares". The equivalent procedure (in this case) of finding parameter values that maximize $L(B_1, B_2, \omega, \sigma)$ is called "maximum likelihood". The maximum likelihood procedure is more general than least squares: it has theoretical justification when the likelihood is not Gaussian. The departure of Jaynes was to use (2.2) in Bayes' theorem (1.3), and then to remove the phase and amplitude from further consideration by integration over these parameters.

In doing this preliminary calculation we will make a number of simplifying assumptions, then in Chapter 3 correct them by solving a much more general problem exactly. For now we insert the model (2.1) into the likelihood (2.2) and expand the exponent to obtain:

$$L(B_1, B_2, \omega, \sigma) \propto \sigma^{-N} \exp\left\{-\frac{NQ}{2\sigma^2}\right\} \tag{2.3}$$

where

$$Q \equiv \overline{d^2} - \frac{2}{N}[B_1 R(\omega) + B_2 I(\omega)] + \frac{1}{2}(B_1^2 + B_2^2),$$

and

$$R(\omega) = \sum_{i=1}^{N} d_i \cos(\omega t_i) \tag{2.4}$$

$$I(\omega) = \sum_{i=1}^{N} d_i \sin(\omega t_i) \tag{2.5}$$

are the functions introduced in (1.1), and

$$\overline{d^2} = \frac{1}{N}\sum_{i=1}^{N} d_i^2$$

is the observed mean-square data value. In this preliminary discussion we assumed the data have zero mean value (any nonzero average value has been subtracted from the data), and we simplified the quadratic term as follows:

$$\sum_{i=1}^{N} f(t_i)^2 = B_1^2 \sum_{i=1}^{N} \cos^2 \omega t_i + B_2^2 \sum_{i=1}^{N} \sin^2 \omega t_i + 2B_1 B_2 \sum_{i=1}^{N} \cos(\omega t_i)\sin(\omega t_i),$$

with

$$\sum_{i=1}^{N} \cos^2 \omega t_i = \frac{N}{2} + \frac{1}{2}\sum_{i=1}^{N} \cos 2\omega t_i \simeq \frac{N}{2},$$

$$\sum_{i=1}^{N} \sin^2 \omega t_i = \frac{N}{2} - \frac{1}{2}\sum_{i=1}^{N} \cos 2\omega t_i \simeq \frac{N}{2},$$

$$\sum_{i=1}^{N} \cos(\omega t_i)\sin(\omega t_i) = \frac{1}{2}\sum_{i=1}^{N} \sin(2\omega t_i) \ll \frac{N}{2}$$

so the quadratic term is approximately

$$\sum_{i=1}^{N} f(t_i)^2 \approx \frac{N}{2}\left(B_1^2 + B_2^2\right).$$

The neglected terms are of order one, and small provided $N \gg 1$ (except in the special case $\omega t_N \ll 1$). We will assume, for now, that the data contain no evidence of a low frequency.

The cross term, $\sum_{i=1}^{N} \cos(\omega t_i)\sin(\omega t_i)$, is at most of the same order as the terms we just ignored; therefore, this term is also ignored. The assumption that this cross term is zero is equivalent to assuming the sine and cosine functions are orthogonal on the discrete time sampled set. Indeed, this is the actual case for uniformly spaced time intervals; however, even without uniform spacing this is a good approximation provided N is large. The assumption that the cross terms are zero by orthogonality will prove to be the key to generalizing this problem to more complex models, and in Chapter 3 the assumptions that we are making now will become exact by a change of variables.

2.3 Elimination of Nuisance Parameters

In a harmonic analysis one is primarily interested in the frequency ω. Then if the amplitude, phase, and the variance of the noise are unknown, they are nuisance parameters. We gave the general procedure for dealing with nuisance parameters in Chapter 1. To apply that rule we must integrate the posterior probability density with respect to B_1, B_2, and also σ if the noise variance is unknown.

If we had prior information about the nuisance parameters (such as: they had to be positive, they could not exceed an upper limit, or we had independently measured values for them) then here would be the place to incorporate that information into the calculation. We illustrate the effects of integrating over a nuisance parameter, as well as the use of prior information in, Appendices B and C and explicitly calculate the expectation values of B_1 and B_2 when a prior measurement is available. At present we assume no prior information about the amplitudes B_1 and B_2 and assign them a prior probability which indicates "complete ignorance of a location parameter". This prior is a uniform, flat, prior density; it is called an improper prior probability because it is not normalizable. In principle, we should approach an improper prior as the limit of a sequence of proper priors. The point is discussed further in Appendices A and B. However, in this problem there are no difficulties with the use of the uniform prior because the Gaussian cutoff in the likelihood function ensures convergence in (2.3).

Upon multiplying the likelihood (2.3) by the uniform prior and integrating with respect to B_1 and B_2 one obtains the joint quasi-likelihood of ω and σ:

$$L(\omega,\sigma) \propto \sigma^{-N+2} \times \exp\left\{-\frac{N}{2\sigma^2}[\overline{d^2} - 2C(\omega)/N]\right\} \tag{2.6}$$

where $C(\omega)$, the Schuster periodogram defined in (1.1), has appeared in a very natural and unavoidable way. If one knows the variance σ^2 from some independent source and has no additional prior information about ω, then the problem is completed. The posterior probability density for ω is given by

$$P(\omega|D,\sigma,I) \propto \exp\left\{\frac{C(\omega)}{\sigma^2}\right\}. \tag{2.7}$$

Because we have assumed little prior information about B_1, B_2, ω and have made conservative assumptions about the noise; this probability density will yield conservative estimates of ω. By this we mean, as before, that if we had more prior information, we could exploit it to obtain still better results. We will illustrate this point further

in Chapter 5 and show that when the data have characteristics which differ from our assumptions, Eq. (2.7) will always make a conservative estimates of the frequency ω. Thus the assumptions we are making act as safeguards to protect us from seeing things in the data that are not really there. We place such great stress on this point because we shall presently obtain some surprisingly sharp estimates.

The above analysis is valid whenever the noise variance (or power) is known. Frequently one has no independent prior knowledge of the noise. The noise variance σ^2 then becomes a nuisance parameter. We eliminate it in much the same way as the amplitudes were eliminated. Now σ is restricted to positive values and additionally it is a scale parameter. The prior which indicates "complete ignorance" of a scale is the Jeffreys prior $1/\sigma$ [15]. Multiplying Eq. (2.6) by the Jeffreys prior and integrating over all positive values gives

$$P(\omega|D,I) \propto \left[1 - \frac{2C(\omega)}{N\overline{d^2}}\right]^{\frac{2-N}{2}}. \tag{2.8}$$

This is called a "Student t-distribution" for historical reasons, although it is expressed here in very nonstandard notation. In our case it is the posterior probability density that a stationary harmonic frequency ω is present in the data when we have no prior information about σ.

These simple results, Eqs. (2.7) and (2.8), show why the discrete Fourier transform tends to peak at the location of a frequency when the data are noisy. Namely, the discrete Fourier transform is directly related to the probability that a single harmonic frequency is present in the data, even when the noise level is unknown. Additionally, zero padding a time series (i.e. adding zeros at its end to make a longer series) and then taking the Fast Fourier transform of the padded series, is equivalent to calculating the Schuster periodogram at smaller frequency intervals. If the signal one is analyzing is a simple harmonic frequency plus noise, then the maximum of the periodogram will be the "best" estimate of the frequency that we can make in the absence of additional prior information about it.

We now see the discrete Fourier transform and the Schuster periodogram in a entirely new light: the highest peak in the discrete Fourier transform is an optimal frequency estimator for a data set which contains a single harmonic frequency in the presence of Gaussian white noise. Stated more carefully, the discrete Fourier

transform will give optimal frequency estimates if six conditions are met:

1. The number of data values N is large,
2. There is no constant component in the data,
3. There is no evidence of a low frequency,
4. The data contain only one frequency,
5. The frequency must be stationary
   (i.e. the amplitude and phase are constant),
6. The noise is white.

If any of these six conditions is not met, the discrete Fourier transform may give misleading or simply incorrect results in light of the more realistic models. Not because the discrete Fourier transform is wrong, but because it is answering what we should regard as the wrong question. The discrete Fourier transform will always interpret the data in terms of a single harmonic frequency model! In Chapter 6 we illustrate the effects of violating one or more of these assumptions and demonstrate that when they are violated the estimated parameters are always less certain than when these conditions are met.

2.4 Resolving Power

When the six conditions are met, just how accurately can the frequency be estimated? This question is easily answered; we do this by approximating (2.7) by a Gaussian and then making the (mean) $\pm$ (standard deviation) estimates of the frequency ω. Expanding $C(\omega)$ about the maximum $\hat{\omega}$ we have

$$C(\omega) = C(\hat{\omega}) - \frac{b}{2}(\hat{\omega} - \omega)^2 + \cdots$$

where

$$b \equiv -C''(\hat{\omega}) > 0. \tag{2.9}$$

The Gaussian approximation is

$$P(\omega|D,\sigma,I) \simeq \left[\frac{2b}{\pi\sigma^2}\right]^{\frac{1}{2}} \exp\left\{-\frac{b(\hat{\omega}-\omega)^2}{2\sigma^2}\right\}$$

from which we would make the (mean) $\pm$ (standard deviation) estimate of the frequency

$$\omega_{\text{est}} = \hat{\omega} \pm \frac{\sigma}{\sqrt{b}}.$$

The accuracy depends on the curvature of $C(\omega)$ at its peak, not on the height of $C(\omega)$. For example, if the data are composed of a single sine wave plus noise $e(t)$ of standard deviation σ

$$d_t = \hat{B}_1 \cos(\hat{\omega} t) + e_t$$

then as found by Jaynes [12]:

$$C(\omega_{\max}) \simeq \frac{N\hat{B}_1^2}{4}$$

$$b \simeq \frac{\hat{B}_1^2 N^3}{48}$$

$$(\omega)_{\text{est}} = \hat{\omega} \pm \frac{\sigma}{|B_1|}\sqrt{48/N^3} \tag{2.10}$$

which indicates, as intuition would lead us to expect, that the accuracy depends on the signal-to-noise ratio, and quite strongly on how much data we have.

The height of the posterior probability density increases like the exponential of $N\hat{B}_1^2/4\sigma^2$ while the error estimates depend on the exponential of $N^3\hat{B}_1^2/96\sigma^2$. If one has a choice between doubling the amount of data $N \to 2N$, or doubling the signal-to-noise ratio $\hat{B}_1/\sigma \to 2\hat{B}_1/\sigma$, always double the amount of data if you have detected the signal, and always double the signal-to-noise ratio if you have no strong evidence of a signal.

If we have sufficient signal-to-noise ratio for the posterior probability density $\exp\{N\hat{B}_1^2/4\sigma^2\}$ to have a peak well above the noise, doubling the amount of data, $N \to 2N$ will double the height of the periodogram giving $\exp\{N\hat{B}_1^2/4\sigma^2\}$ times more evidence of a frequency while the error will go down like $\sqrt{8}$. On the other hand, if the signal-to-noise ratio is so low that $\exp\{N\hat{B}_1^2/4\sigma^2\}$ has no clear peak above the noise, then doubling the signal-to-noise ratio $\hat{B}_1^2/\sigma^2 \to 4\hat{B}_1^2/\sigma^2$ will give $\exp\{3N\hat{B}_1^2/4\sigma^2\}$ times more evidence for a frequency, while the error goes down by 2. The trade off is clear: if you have sufficient signal-to-noise for signal detection more data are important for resolution; otherwise more signal-to-noise will detect the signal with less data.

We can further compare these results with experience, but first note that we are using dimensionless units, since we took the data sampling interval to be 1. Converting to ordinary physical units, let the sampling interval be Δt seconds, and denote by f the frequency in Hz. Then the total number of cycles in our data record is

$$\frac{\hat{\omega}(N-1)}{2\pi} = (N-1)\hat{f}\Delta t = \hat{f}T$$

where $T = (N-1)\Delta t$ seconds is the duration of our data run. So the conversion of dimensionless ω to f in physical units is

$$f = \frac{\omega}{2\pi\Delta t} \text{ Hz.}$$

The frequency estimate (2.10) becomes

$$f_{\text{est}} = \hat{f} \pm \delta f \text{ Hz}$$

where now, not distinguishing between N and $(N-1)$,

$$\delta f = \frac{\sigma}{2\pi\hat{B}_1 T}\sqrt{48/N} = \frac{1.1\sigma}{\hat{B}_1 T\sqrt{N}} \text{ Hz.} \tag{2.11}$$

Comparing this with (2.10) we now see that to improve the accuracy of the estimate the two most important factors are how long we sample (the T dependence) and the signal-to-noise ratio. We could double the number of data values in one of two ways, by doubling the total sampling time or by doubling the sampling rate. However, (2.11) clearly indicates that doubling the sampling time is to be preferred. This indicates that data values near the beginning and end of a record are most important for frequency estimation, in agreement with intuitive common sense.

Let us take a specific example: if we have an RMS signal-to-noise ratio (i.e. ratio of RMS signal to RMS noise $\equiv$ S/N) of S/N $= \hat{B}_1/\sqrt{2}\sigma = 1$, and we take data every $\Delta t = 10^{-3}$ sec. for $T = 1$ second, thus getting $N = 1000$ data points, the theoretical accuracy for determining the frequency of a single steady sinusoid is

$$\delta f = \frac{1.1}{\sqrt{2000}} = 0.025 \text{ Hz} \tag{2.12}$$

while the Nyquist frequency for the onset of aliasing is $f_N = (2\Delta t)^{-1} = 500$Hz, greater by a factor of 20,000.

To some, this result will be quite startling. Indeed, had we considered the periodogram itself to be a spectrum estimator, we would have calculated instead the width of its central peak. A noiseless sinusoid of frequency $\hat{\omega}$ would have a periodogram proportional to

$$C(\omega) \propto \frac{\sin^2\{N(\omega-\hat{\omega})/2\}}{\sin^2\{(\omega-\hat{\omega})/2\}}$$

thus the half-width at half amplitude is given by $|N(\hat{\omega}-\omega)/2| = \pi/4$ or $\delta\omega = \pi/2N$. Converting to physical units, the periodogram will have a width of about

$$\delta f = \frac{1}{4N\Delta t} = \frac{1}{4T} = 0.25 \text{ Hz} \tag{2.13}$$

just ten times greater than the value (2.12) indicated by probability theory. This factor of ten is the amount of narrowing produced by the exponential peaking of the periodogram in (2.7), even for unit signal-to-noise ratio.

But some would consider even the result (2.13) to be a little overoptimistic. The famous Rayleigh criterion [16] for resolving power of an optical instrument supposes that the minimum resolvable frequency difference corresponds to the peak of the periodogram of one sinusoid coming at the first zero of the periodogram of the second. This is twice (2.13):

$$\delta f_{\text{Rayleigh}} = \frac{1}{2T} = 0.5 \text{ Hz}. \tag{2.14}$$

There is a widely believed "folk-theorem" among theoreticians without laboratory experience, which seems to confuse the Rayleigh limit with the Heisenberg uncertainty principle, and holds that (2.14) is a fundamental irreducible limit of resolution. Of course there is no such theorem, and workers in high resolution NMR have been routinely determining line positions to an accuracy that surpasses the Rayleigh limit by an order of magnitude, for thirty years.

The misconception is perhaps strengthened by the curious coincidence that (2.14) is also the minimum half-width that can be achieved by a Blackman-Tukey spectrum analysis [13] (even at infinite signal-to-noise ratio) because the "Hanning window" tapering function that is applied to the data to suppress side-lobes (the secondary maxima of $[\sin(x)/x]^2$) just doubles the width of the periodogram. Since the Blackman-Tukey method has been used widely by economists, oceanographers, geophysicists, and engineers for many years, it has taken on the appearance of an optimum procedure.

According to E.T. Jaynes, Tukey himself acknowledged [17] that his method fails to give optimum resolution, but held this to be of no importance because "real time series do not have sharp lines." Nevertheless, this misconception is so strongly held that there have been attacks on the claims of Bayesian/Maximum Entropy spectrum analysts to be able to achieve results like (2.12) when the assumed conditions are met. Some have tried to put such results in the same category with circle squaring and perpetual motion machines. Therefore we want to digress to explain in very elementary physical terms why it is the Bayesian result (2.11) that does correspond to what a skilled experimentalist can achieve.

Suppose first that our only data analysis tool is our own eyes looking at a plot of the raw data of duration $T = 1$ sec., and that the unknown frequency f in (2.12) is 100Hz. Now anyone who has looked at a record of a sinusoid and equal amplitude

wide-band noise, knows that the cycles are quite visible to the eye. One can count the total number of cycles in the record confidently (using interpolation to help us over the doubtful regions) and will feel quite sure that the count is not in error by even one cycle. Therefore by raw eyeballing of the data and counting the cycles, one can achieve an accuracy of

$$\delta f \simeq \frac{1}{T} = 1 \text{ Hz}.$$

But in fact, if one draws the sine wave that seems to fit the data best, he can make a quite reliable estimate of how many quarter-cycles were in the data, and thus achieve

$$\delta f \simeq \frac{1}{4T} = 0.25 \text{ Hz}$$

corresponding just to the periodogram width (2.13).

Then the use of probability theory needs to surpass the naked eye by another factor of ten to achieve the Bayesian width (2.12). What probability theory does is essentially to average out the noise in a way that the naked eye cannot do. If we repeat some measurement N times, any randomly varying component of the data will be suppressed relative to the systematic component by a factor of $N^{-\frac{1}{2}}$, the standard rule.

In the case considered, we assumed $N = 1000$ data points. If they were all independent measurements of the same quantity with the same accuracy, this would suppress the noise by about a factor of 30. But in our case not all measurements are equally cogent for estimating the frequency. Data points in the middle of the record contribute very little to the result; only data points near the ends are highly relevant for determining the frequency, so the effective number of observations is less than 1000. The probability analysis leading to (2.12) indicates that the "effective number of observations" is only about $N/10 = 100$; thus the Bayesian width (2.12) that results from the exponential peaking of the periodogram now appears to be, if anything, somewhat conservative.

Indeed, that is what Bayesian analysis always does when we use smooth, uninformative priors for the parameters, because then probability theory makes allowance for all possible values that they might have. As noted before, if we had any cogent prior information about ω and expressed it in a narrower prior, we would be led to still better results; but they would not be much better unless the prior range became comparable to the width of the likelihood $L(\omega)$.

2.5 The Power Spectral Density $\hat{p}$

The usual way the result from a spectral analysis is displayed is in the form of a power spectral density (i.e. power per unit frequency). In Fourier transform spectroscopy this is typically taken as the squared magnitude of the discrete Fourier transform of the data. We would like to express the results of the present calculation in a similar manner to facilitate comparisons between these techniques, although strictly speaking there is no exact correspondence between a spectral density defined with reference to a stochastic model and one that pertains to a parameter estimation model.

We begin by defining what we mean by the "estimated spectrum," since several quite different meanings of the term can be found in the literature. Define $\hat{p}(\omega)d\omega$ as the expectation, over the joint posterior probability distribution for all the parameters, of the energy carried by the signal (not the noise) in frequency range $d\omega$, during our observation time $t_N - t_1$. Then $\int \hat{p}(\omega)d\omega$ over some frequency range is the expectation of the total energy carried by the signal in that frequency range. The total energy E carried by the signal in our model is

$$E = \int_{t_1}^{t_N} f(t)^2 dt \approx \frac{T}{2}\left(B_1^2 + B_2^2\right)$$

and its expectation is given by

$$\hat{p}(\omega) = \frac{T}{2}\langle B_1^2 + B_2^2\rangle;$$

but $N = T/\Delta t$, where Δt is the sampling time which in dimensionless units is one. The power spectral density is

$$\hat{p}(\omega) = \frac{N}{2}\int dB_1 dB_2 (B_1^2 + B_2^2) P(\omega, B_1, B_2|D, \sigma, I).$$

Performing the integrals over B_1 and B_2 we obtain

$$\hat{p}(\omega) = 2\left[\sigma^2 + C(\omega)\right] P(\omega|D, \sigma, I). \tag{2.15}$$

We see now that the peak of the periodogram is indicative of the total energy carried by the signal. The additional term $2\sigma^2$ is not difficult to explain; but we delay that explanation until after we have derived these results for the general theory (see page 52).

If the noise variance is assumed known, (2.15) becomes

$$\hat{p}(\omega) = 2\left[\sigma^2 + C(\omega)\right] \frac{\exp\left\{C(\omega)/\sigma^2\right\}}{\int d\omega \exp\left\{C(\omega)/\sigma^2\right\}}. \tag{2.16}$$

Probability theory will handle those secondary maxima (side lobes) that occur in the periodogram by assigning them negligible weight.

This is easily seen by considering the same example discussed earlier. Take $d(t) = \hat{B}_1 \cos(\hat{\omega}t)$ sampled on a uniform grid; then when $\hat{\omega} \simeq \omega$

$$C(\omega) \simeq \frac{\hat{B}_1^2}{4N}\left[\frac{\sin N(\hat{\omega}-\omega)/2}{(\hat{\omega}-\omega)/2}\right]^2$$

and C'' is

$$C'' \equiv b \simeq \frac{\hat{B}_1^2 N^3}{24}$$

and $\hat{p}(\omega)$ is approximately

$$\hat{p}(\omega) \simeq 2\left[\sigma^2 + \frac{4\hat{B}_1^2 \sin^2 N(\hat{\omega}-\omega)/2}{(\hat{\omega}-\omega)^2}\right]\left[\frac{\hat{B}_1^2 N^3}{24\pi\sigma^2}\right]^{\frac{1}{2}} \exp\left\{-\frac{\hat{B}_1^2 N^3}{48\sigma^2}(\hat{\omega}-\omega)^2\right\}.$$

Unless the signal-to-noise ratio $\hat{B}_1/\sigma\sqrt{2}$ is very small, this is very nearly a delta function.

If we take $\hat{B}_1 = \sqrt{2}\sigma = 1$, and $N = 1000$ data values, then

$$\hat{p}(\omega) \simeq 2\left[1 + 4\frac{\sin^2 1000(\hat{\omega}-\omega)/2}{(\hat{\omega}-\omega)^2}\right][5150]\exp\{-4\times 10^7(\hat{\omega}-\omega)^2\}.$$

This reaches a maximum value of 10^{11} at $\hat{\omega} = \omega$ and has dropped to $\frac{1}{2}$ this value when $\hat{\omega} - \omega$ has changed by only 0.0001; this function is indeed a good approximation to a delta function and (2.16) may be approximated by:

$$\hat{p}(\omega) \simeq \left[\sigma^2 + C(\hat{\omega})\right][\delta(\hat{\omega}-\omega) + \delta(\hat{\omega}+\omega)]$$

for most purposes. But for the term σ^2, the peak of the periodogram is, in our model, nearly the total energy carried by the signal. It is not an indication of the spectral density as Schuster [11] supposed it to be for a stochastic model. In the present model, the periodogram of the data is not even approximately the spectral energy density of the signal.

2.6 Wolf's Relative Sunspot Numbers

Wolf's relative sunspot numbers are, perhaps, the most analyzed set of data in all of spectrum analysis. As Marple [3] explains in more detail, these numbers (defined as: $W = k[10g + f]$, where g is the number of sunspot groups, f is the number of individual sunspots, and k is used to reduce different telescopes to a common scale) have been collected on a yearly basis since 1700, and on a monthly basis since 1748 [18]. The exact physical mechanism which generates the sunspots is unknown, and no complete theory exists. Different analyses of these numbers have been published more or less regularly since their tabulation began. Here we will analyze the sunspot numbers with a number of different models including the simple harmonic analysis just completed, even though we know this analysis is too simple to be realistic for these numbers.

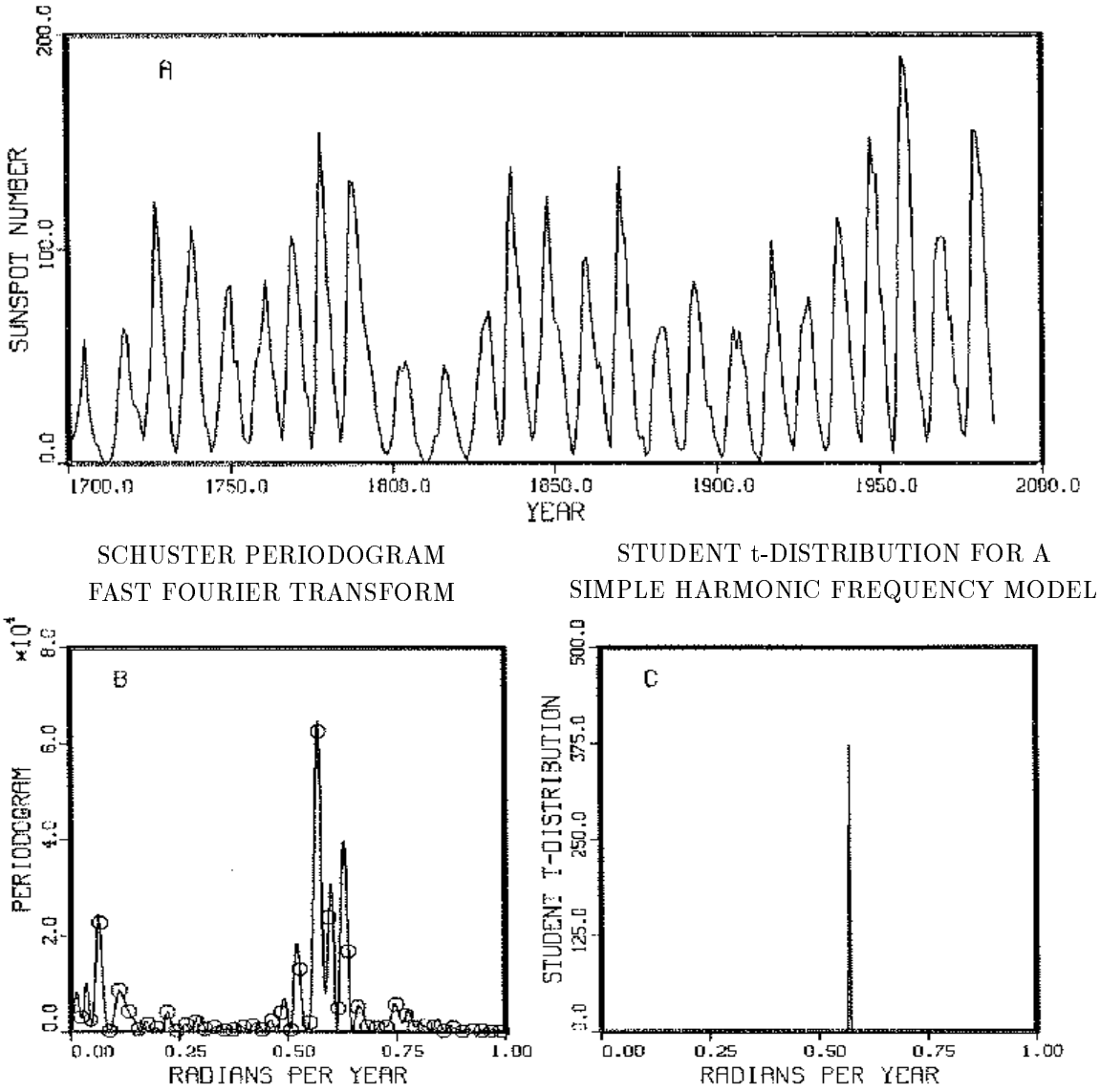
We have plotted the time series from 1700 to 1985 in Fig. 2.1(A). A cursory examination of this time series does indeed show a cyclic variation with a period of about 11 years. The square of the discrete Fourier transform is a continuous function of frequency and is proportional to the Schuster periodogram of the data Fig. 2.1(B), continuous curve. The frequencies could be restricted to the Nyquist [19] [20] steps (open circles); it is a theorem that the discrete Fourier transform on those points contains all the information that is in the periodogram, but one sees that the information is much more apparent to the eye in the continuous periodogram. The Schuster periodogram or the discrete Fourier transform clearly show a maximum with period near 11 years.

We then compute the "Student t-distribution" (2.8) and have displayed it in figure 2.1(C). Now because of the processing in (2.8) all details in the periodogram have been suppressed and only the peak at 11 years remains.

We determine the accuracy of the frequency estimate as follows: we locate the maximum of the "Student t-distribution", integrate about a symmetric interval, and record the enclosed probability at a number of points. This gives a period of 11.04 years with

period in years		accuracy in years	probability enclosed
11.04	$\pm$	0.015	0.62
	$\pm$	0.020	0.75
	$\pm$	0.026	0.90

Figure 2.1: Wolf's Relative Sunspot Numbers



Wolf's relative sunspot numbers (A) have been collected on a yearly basis since 1700. The periodogram (B) contains evidence of several complex phenomena. In spite of this the single frequency model posterior probability density (C) picks out the 11.04 year cycle to an estimated accuracy of ± 10 days.

as an indication of the accuracy. According to this, there is not one chance in 10 that the true period differs from 11.04 years by more than 10 days. At first glance, this appears too good to be true.

But what does raw eyeballing of the data give? In 285 years, there are about $285/11 \approx 26$ cycles. If we can count these to an accuracy of $\pm 1/4$ cycle, our period estimate would be about

$$(f)_{\rm est} = 11 \text{ years} \pm 39 \text{ days}.$$

Probability averaging of the noise, as discussed above (2.10), would reduce this uncertainty by about a factor of $\sqrt{285/10} = 5.3$, giving

$$(f)_{\rm est} = 11 \text{ years} \pm 7.3 \text{ days}, \quad \text{or} \quad (f)_{\rm est} = 11 \pm 0.02 \text{ years}$$

which corresponds nicely with the result of the probability analysis.

These results come from analyzing the data by a model which said there is nothing present but a single sinusoid plus noise. Probability theory, given this model, is obliged to consider everything in the data that cannot be fit to a single sinusoid to be noise. But a glance at the data shows clearly that there is more present than our model assumed: therefore, probability theory must estimate the noise to be quite large.

This suggests that we might do better by using a more realistic model which allows the "signal" to have more structure. Such a model can be fit to the data more accurately; therefore it will estimate the noise to be smaller. This should permit a still better period estimate!

But caution forces itself upon us; by adding more and more components to the model we can always fit the data more and more accurately; it is absurd to suppose that by mere proliferation of a model we can extract arbitrarily accurate estimates of a parameter. There must be a point of diminishing returns – or indeed of negative returns – beyond which we are deceiving ourselves.

It is very important to understand the following point. Probability theory always gives us the estimates that are justified by the information *that was actually used* in the calculation. Generally, a person who has more relevant information will be able to do a different (more complicated) calculation, leading to better estimates. But of course, this presupposes that the extra information is actually true. If one puts false information into a probability calculation, then probability theory will give optimal estimates based on false information: these could be very misleading. The onus is

always on the user to tell the truth and nothing but the truth; probability theory has no safety device to detect falsehoods.

The issue just raised takes us into an area that has been heretofore, to the best of our knowledge, unexplored by any coherent theory. The analysis of this section has shown how to make the optimum estimates of parameters *given a model* whose correctness is not questioned. Deeper probability analysis is needed to indicate how to make the optimum choice of a model, which neither cheats us by giving poorer estimates than the data could justify, nor deceives us by seeming to give better estimates than the data can justify. But before we can turn to the model selection problem, the results of this chapter must be generalized to more complex models and it is to this task that we now turn.

Chapter 3

THE GENERAL MODEL EQUATION PLUS NOISE

The results of the previous chapter already represent progress on the spectral analysis problem because we were able to remove consideration of the amplitude, phase and noise level, and find what probability theory has to say about the frequency alone. In addition, it has given us an indication about how to proceed to more general problems. If we had used a model where the quadratic term in the likelihood function did not simplify, we would have a more complicated analytical solution. Although any multivariate Gaussian integral can be done, the key to being able to remove the nuisance parameters easily, and above all selectively, was that the likelihood factored into independent parts. In the full spectrum analysis problem worked on by Jaynes, [12] the nuisance parameters were not independent, and the explicit solution required the diagonalization of a matrix that could be quite large.

3.1 The Likelihood Function

To understand an easier approach to complex models, suppose we have a model of the form

$$d_i = f(t_i) + e_i$$
$$f(t) = \sum_{j=1}^{m} B_j G_j(t, \{\omega\}). \tag{3.1}$$

The model functions, $G_i(t, \{\omega\})$, are themselves functions of a set of parameters which we label collectively $\{\omega\}$ (these parameters might be frequencies, chirp rates, decay

rates, or any other quantities one could encounter). Now if we substitute this model into the likelihood (2.2), the simplification that occurred in (2.3) does not take place:

$$L(\{B\},\{\omega\},\sigma) \propto \sigma^{-N} \times \exp\{-\frac{NQ}{2\sigma^2}\} \tag{3.2}$$

where

$$Q \equiv \overline{d^2} - \frac{2}{N}\sum_{j=1}^{m}\sum_{i=1}^{N} B_j d_i G_j(t_i) + \frac{1}{N}\sum_{j=1}^{m}\sum_{k=1}^{m} g_{jk} B_j B_k \tag{3.3}$$

$$g_{jk} = \sum_{i=1}^{N} G_j(t_i) G_k(t_i). \tag{3.4}$$

If the desired simplification is to take place, the matrix g_{jk} must be diagonal.

3.2 The Orthonormal Model Equations

For the matrix g_{jk} to be diagonal the model functions G_j must be made orthogonal. This can be done by taking appropriate linear combinations of them. But care must be taken; we do not desire a set of orthogonal functions of a continuous variable t, but a set of vectors which are orthogonal when summed over the discrete sampling times t_i. It is the sum over t_i appearing in the quadratic term of the likelihood which must simplify.

To accomplish this, consider the real symmetric matrix g_{jk} (3.4) defined above. Since for all x_j satisfying $\sum x_j^2 > 0$,

$$\sum_{j,k=1}^{m} g_{jk} x_j x_k = \sum_{i=1}^{N} \left(\sum_{j=1}^{m} x_j G_j(t_i) \right)^2 \geq 0$$

so that g_{jk} is positive definite if it is of rank m. If it is of rank $r < m$, then the model functions $G_j(t)$ or the sampling times t_i were poorly chosen. That is, if a linear combination of the $G_j(t)$ is zero at every sampling point:

$$\sum_{j=1}^{m} x_j G_j(t_i) = 0, \qquad (1 \leq i \leq N)$$

then at least one of the model functions $G_j(t)$ is redundant and can be removed from the model without changing the problem.

We suppose that redundant model functions have been removed, so that g_{jk} is positive definite and of rank m in what follows. Let e_{kj} represent the jth component

of the kth normalized eigenvector of g_{jk}; i.e.

$$\sum_{k=1}^{m} g_{jk} e_{lk} = \lambda_l e_{lj},$$

where λ_l is the lth eigenvalue of g_{jk}. Then the functions $H_j(t)$, defined as

$$H_j(t) = \frac{1}{\sqrt{\lambda_j}} \sum_{k=1}^{m} e_{jk} G_k(t), \tag{3.5}$$

have the desired orthonormality condition,

$$\sum_{i=1}^{N} H_j(t_i) H_k(t_i) = \delta_{jk}. \tag{3.6}$$

The model Eq. (3.1) can now be rewritten in terms of these orthonormal functions as

$$f(t) = \sum_{k=1}^{m} A_k H_k(t). \tag{3.7}$$

The amplitudes B_k are linearly related to the A_k by

$$B_k = \sum_{j=1}^{m} \frac{A_j e_{jk}}{\sqrt{\lambda_j}} \quad \text{and} \quad A_k = \sqrt{\lambda_k} \sum_{j=1}^{m} B_j e_{kj}. \tag{3.8}$$

The volume elements are given by

$$\begin{aligned} dB_1 \cdots dB_m d\omega_1 \cdots d\omega_r &= \left| \frac{\mathrm{e}_{lj}}{\sqrt{\lambda_j}} \right| dA_1 \cdots dA_m d\omega_1 \cdots d\omega_r \\ &= \lambda_1^{-\frac{1}{2}} \cdots \lambda_m^{-\frac{1}{2}} dA_1 \cdots dA_m d\omega_1 \cdots d\omega_r. \end{aligned} \tag{3.9}$$

The Jacobian is a function of the $\{\omega\}$ parameters and is a constant as long as we are not integrating over these $\{\omega\}$ parameters. At the end of the calculation the linear relations between the A's and B's can be used to calculate the expected values of the B's from the expected value of the A's, and the same is true of the second posterior moments:

$$E(B_k|\{\omega\}, D, I) = \langle B_k \rangle = \sum_{j=1}^{m} \frac{\langle A_j \rangle e_{jk}}{\sqrt{\lambda_j}} \tag{3.10}$$

$$E(B_k B_l|\{\omega\}, D, I) = \langle B_k B_l \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{e_{ik} e_{jl} \langle A_i A_j \rangle}{\sqrt{\lambda_i \lambda_j}} \tag{3.11}$$

where $E(B_k|D, I)$ stands for the expectation value of B_k given the data D, and the prior information I: this is the notation used by the general statistical community, while $\langle B_k \rangle$ is the notation more familiar in the physical sciences.

The two operations of making a transformation on the model functions and changing variables will transform any nonorthonormal model of the form (3.1) into an orthonormal model (3.7). We still have a matrix to diagonalize, but this is done once at the beginning of the calculation. If the g_{jk} matrix cannot be diagonalized analytically, it can still be computed numerically and then diagonalized. It is not necessary to carry out the inverse transformation if we are interested only in estimating the $\{\omega\}$ parameters: the $H_j(t, \{\omega\})$ are functions of them.

3.3 Elimination of the Nuisance Parameters

We are now in a position to proceed as before. Because the calculation is essentially identical to the single harmonic frequency calculation we will proceed very rapidly. The likelihood can now be factored into a set of independent likelihoods for each of the A_j. It is now possible to remove the nuisance parameters easily. Using the joint likelihood (3.2), we make the change of function (3.5) and the change of variables (3.8) to obtain the joint likelihood of the new parameters

$$L(\{A\}, \{\omega\}, \sigma) \propto \sigma^{-N} \times \exp\left\{-\frac{N}{2\sigma^2}[\overline{d^2} - \frac{2}{N}\sum_{j=1}^{m} A_j h_j + \frac{1}{N}\sum_{j=1}^{m} A_j^2]\right\} \tag{3.12}$$

$$h_j \equiv \sum_{i=1}^{N} d_i H_j(t_i), \qquad (1 \leq j \leq m). \tag{3.13}$$

Here h_j is just the projection of the data onto the orthonormal model function H_j. In the simple harmonic analysis performed in Chapter 2, the $R(\omega)$ and $I(\omega)$ functions are the analogues of these h_j functions. However, the h_j functions are more general, we did not make any approximations in deriving them. The orthonormality of the H_j functions was used to simplify the quadratic term. This simplification makes it possible to complete the square in the likelihood and to integrate over the A_j's, or any selected subset of them.

As before, if one has prior information about these amplitudes, then here is where it should be incorporated. Because we are performing a general calculation and have not specified the model functions we assume no prior information is available about the amplitudes, and thus obtain conservative estimates by assigning the amplitudes a uniform prior. Performing the m integrations one obtains

$$L(\{\omega\}, \sigma) \propto \sigma^{-N+m} \times \exp\left\{-\frac{N\overline{d^2} - m\overline{h^2}}{2\sigma^2}\right\} \tag{3.14}$$

where

$$\overline{h^2} \equiv \frac{1}{m}\sum_{j=1}^{m} h_j^2 \tag{3.15}$$

is the mean-square of the observed projections. This equation is the analogue of (2.6) in the simple harmonic calculation. Although it is exact and far more general, it is actually simpler in structure and gives us a better intuitive understanding of the problem than (2.6), as we will see in the Bessel inequality below. In a sense $\overline{h^2}$ is a generalization of the periodogram to arbitrary model functions. In its dependence on the parameters $\{\omega\}$ it is a sufficient statistic for all of them.

If σ is known, then the problem is again completed, provided we have no additional prior information. The joint posterior probability of the $\{\omega\}$ parameters, conditional on the data and our knowledge of σ, is

$$P(\{\omega\}|D,\sigma,I) \propto \exp\left\{\frac{m\overline{h^2}}{2\sigma^2}\right\}. \tag{3.16}$$

But if there is no prior information available about the noise, then σ is a nuisance parameter and can be eliminated as before. Using the Jeffreys prior $1/\sigma$ and integrating (3.14) over σ gives

$$P(\{\omega\}|D,I) \propto \left[1 - \frac{m\overline{h^2}}{N\overline{d^2}}\right]^{\frac{m-N}{2}}. \tag{3.17}$$

This is again of the general form of the "Student t-distribution" that we found before in (2.8). But one may be troubled by the negative sign [in the square brackets (3.17)], which suggests that (3.17) might become singular. We pause to investigate this possibility by Bessel's famous argument.

3.4 The Bessel Inequality

Suppose we wish to approximate the data vector $\{d_1,\cdots,d_N\}$ by the orthogonal functions $H_j(t)$:

$$d_i = \sum_{j=1}^{m} a_j H_j(t_i) + \text{error}, \qquad (1 \le i \le N).$$

What choice of $\{a_1,\cdots,a_m\}$ is "best"? If our criterion of "best" is the mean-square error, we have

$$\begin{aligned} 0 &\leq \sum_{i=1}^{N}\left[d_i - \sum_{j=1}^{m} a_j H_j(t_i)\right]^2 \\ &= N\overline{d^2} + \sum_{j=1}^{m}(a_j^2 - 2a_j h_j) \\ &= N\overline{d^2} - m\overline{h^2} + \sum_{j=1}^{m}(a_j - h_j)^2 \end{aligned}$$

where we have used (3.13) and the orthonormality (3.6). Evidently, the "best" choice of the coefficients is

$$a_j = h_j, \qquad (1 \leq j \leq m)$$

and with this choice the minimum possible mean-square error is given by the Bessel inequality

$$N\overline{d^2} - m\overline{h^2} \geq 0 \tag{3.18}$$

with equality if and only if the approximation is perfect. In other words, Eq. (3.17) becomes singular somewhere in the parameter space if and only if the model

$$f(t) = \sum_{j=1}^{m} A_j H_j(t)$$

can be fitted to the data exactly. But in that case we know the parameters by deductive reasoning, and probability theory becomes superfluous. Even so, probability theory is still working correctly, indicating an infinitely greater probability for the true parameter values than for any others.

3.5 An Intuitive Picture

The Bessel inequality gives us the following intuitive picture of the meaning of Eqs. (3.12-3.17). The data $\{d_1, \cdots, d_N\}$ comprise a vector in an N-dimensional linear vector space S_N. The model equation

$$d_i = \sum_{j=1}^{m} A_j H_j(t_i) + e_i, \qquad (1 \leq i \leq N)$$

supposes that these data can be separated into a "systematic part" $f(t_i)$ and a "random part" e_i. Estimating the parameters of interest $\{\omega\}$ that are hidden in the model

functions $H_j(t)$ amounts essentially to finding the values of the $\{\omega\}$ that permit $f(t)$ to make the closest possible fit to the data by the mean-square criterion. Put differently, probability theory tells us that the most likely values of the $\{\omega\}$ are those that allow a maximum amount of the mean-square data $\overline{d^2}$ to be accounted for by the systematic term; from (3.18), those are the values that maximize $\overline{h^2}$.

However, we have N data points and only m model functions to fit to them. Therefore, to assign a particular model is equivalent to supposing that the systematic component of the data lies only in an m-dimensional subspace S_m of S_N. What kind of data should we then expect?

Let us look at the problem backwards for a moment. Suppose someone knows (never mind how he could know this) that the model is correct, and he also knows the true values of all the model parameters $(\{A\},\{\omega\},\sigma)$ – call this the Utopian state of knowledge U – but he does not know what data will be found. Then the probability density that he would assign to any particular data set $D=\{d_1,\cdots,d_N\}$ is just our original sampling distribution (3.2):

$$P(D|U)=(2\pi\sigma^2)^{-\frac{N}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[d_i-f(t_i)]^2\right\}.$$

From this he would find the expectations and covariances of the data:

$$\begin{aligned} E(d_i|U)=\langle d_i\rangle &= f(t_i) \qquad (1\le i\le N)\\ \langle d_id_j\rangle-\langle d_i\rangle\langle d_j\rangle &= (2\pi\sigma^2)^{-\frac{N}{2}}\int d^Nx\ x_ix_j\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}x_i^2\right\}\\ &= \sigma^2\delta_{ij}\end{aligned}$$

therefore he would "expect" to see a value of $\overline{d^2}$ of about

$$\begin{aligned} E(\overline{d^2}|U)=\langle\overline{d^2}\rangle &= \frac{1}{N}\sum_{i=1}^{N}\langle d_i^2\rangle\\ &= \frac{1}{N}\sum_{i=1}^{N}(\langle d_i\rangle^2+\sigma^2)\\ &= \frac{1}{N}\sum_{i=1}^{N}f^2(t_i)+\sigma^2.\end{aligned} \tag{3.19}$$

But from the orthonormality (3.6) of the $H_j(t_i)$, we have

$$\begin{aligned}\sum_{i=1}^{N}f^2(t_i) &= \sum_{l=1}^{N}\sum_{j,k=1}^{m}A_jA_kH_j(t_i)H_k(t_i)\\ &= \sum_{j=1}^{m}A_j^2.\end{aligned}$$

So that (3.19) becomes

$$\langle\overline{d^2}\rangle = \frac{m}{N}\overline{A^2} + \sigma^2.$$

Now, what value of $\overline{h^2}$ would he expect the data to generate? This is

$$\begin{aligned} E(\overline{h^2}|U) = \langle\overline{h^2}\rangle &= \frac{1}{m}\sum_{j=1}^{m}\langle h_j^2\rangle \\ &= \frac{1}{m}\sum_{j=1}^{m}\sum_{i,k=1}^{N}\langle d_i d_k\rangle H_j(t_i)H_j(t_k) \qquad (3.20) \\ &= \frac{1}{m}\sum_{j=1}^{m}\sum_{i,k=1}^{N}(\langle d_i\rangle\langle d_k\rangle + \sigma^2\delta_{ik})H_j(t_i)H_j(t_k). \end{aligned}$$

But

$$\begin{aligned} \sum_{i=1}^{N}\langle d_i\rangle H_j(t_i) &= \sum_{i=1}^{N}\sum_{l=1}^{m}A_l H_l(t_i)H_j(t_i) \\ &= \sum_{l=1}^{m}A_l\delta_{lj} \\ &= A_j \end{aligned}$$

and (3.20) reduces to

$$\langle\overline{h^2}\rangle = \overline{A^2} + \sigma^2.$$

So he expects the left-hand side of the Bessel inequality (3.18) to be approximately

$$N\langle\overline{d^2}\rangle - m\overline{h^2} \approx (N-m)\sigma^2. \qquad (3.21)$$

This agrees very nicely with our intuitive judgment that as the number of model functions increases, we should be able to fit the data better and better. Indeed, when $m = N$, the $H_j(t_i)$ become a complete orthonormal set on S_N, and the data can always be fit exactly, as (3.21) suggests.

3.6 A Simple Diagnostic Test

If σ is known, these results give a simple diagnostic test for judging the adequacy of our model. Having taken the data, calculate $(N\overline{d^2} - m\overline{h^2})$. If the result is reasonably close to $(N-m)\sigma^2$, then the validity of the model is "confirmed" (in the sense that the data give no evidence against the model). On the other hand, if $(N\overline{d^2} - m\overline{h^2})$

turns out to be much larger than $(N-m)\sigma^2$, the model is not fitting the data as well as it should: it is "underfitting" the data. That is evidence either that the model is inadequate to represent the data; we could need more model functions, or different model functions, or our supposed value of σ^2 is too low. The next order of business would be to investigate these possibilities.

It is also possible, although unusual, that $(N\overline{d^2} - m\overline{h^2})$ is far less than $(N-m)\sigma^2$; the model is "overfitting" the data. That is evidence either that our supposed value of σ is too large (the data are actually better than we expected), or that the model is more complex than it needs to be. By adding more model functions we can always improve the apparent fit, but if our model functions represent more detail than is really in the systematic effects at work, part of this fit is misleading: we are *fitting the noise.*

A test to confirm this would be to repeat the whole experiment under conditions where we know the parameters should have the same values as before, and compare the parameter estimates from the two experiments. Those parameters that are estimated to be about the same in the two experiments are probably real systematic effects. If some parameters are estimated to be quite different in the two experiments, they are almost surely spurious: i.e. these are not real effects but only artifacts of fitting the noise. The model should then be simplified, by removing the spurious parameters.

Unfortunately, a repetition is seldom possible with geophysical or economic time series, although one may split the data into two parts and see if they make about the same estimates. But repetition is usually easy and standard practice in the controlled environment of a physics experiment. Indeed, the physicist's common-sense criterion of a real effect is its reproducibility. Probability theory does not conflict with good common-sense judgment; it only sharpens it and makes it quantitative. A striking example of this is given in the scenario below.

Consider now the case that σ is completely unknown, where probability theory led us to (3.17). As we show in Appendix C, integrating over a nuisance parameter is very much like estimating the parameter from the data, and then using that estimate in our equations. If the parameter is actually well determined by the data, the two procedures are essentially equivalent. In Chapter 4 we derive an exact expression for

the expectation value of the variance $\langle\sigma^2\rangle$:

$$\begin{aligned}\langle\sigma^2\rangle &= \frac{N}{N-m-2}\left[\overline{d^2}-\frac{m\overline{h^2}}{N}\right] \\ &= \frac{1}{N-m-2}\left[\sum_{i=1}^{N}d_i^2-\sum_{j=1}^{m}h_j^2\right].\end{aligned} \tag{3.22}$$

Constraining σ to this value, we obtain for the posterior probability of the $\{\omega\}$ parameters approximately

$$P(\{\omega\}|D,\langle\sigma^2\rangle,I)\approx\exp\left\{\frac{m\overline{h^2}}{\langle\sigma^2\rangle}\right\}.$$

In effect, probability theory tells us that we should apportion the first m degrees of freedom to the signal, the next two to the variance, and the remaining $(N-m-2)$ should be noise degrees of freedom. Thus everything probability theory cannot fit to the signal will be placed in the noise.

More interesting is the opposite extreme when (3.17) approaches a singular value. Consider the following scenario. You have obtained some data which are recorded automatically to six figures and look like this: $D=\{d_1=1423.16, d_2=1509.77, d_3=1596.38,\cdots\}$. But you have no prior knowledge of the accuracy of those data; for all you know, σ may be as large as 100 or even larger, making the last four digits garbage. But you plot the data, to determine a model function that best fits them. Suppose, for simplicity, that the model function is linear: $d_i=a+si+e_i$. On plotting d_i against i, you are astonished and delighted to see the data falling exactly on a straight line (i.e. to within the six figures given). What conclusions do you draw from this?

Intuitively, one would think that the data must be far "better" than had been thought; you feel sure that $\sigma<10^{-2}$, and that you are therefore able to estimate the slope s to an accuracy considerably better than $\pm10^{-2}$, if the number of data values N is large. It may, however, be hard to see at first glance how probability theory can justify this intuitive conclusion that we draw so easily.

But that is just what (3.17) and (3.22) tell us; Bayesian analysis leads us to it automatically and for any model functions. Even though you had no reason to expect it, if it turns out that the data can be fit almost exactly to a model function, then from the Bessel inequality (3.18) it follows that σ^2 must be extremely small and, if the other parameters are independent, they can all be estimated almost exactly.

By "independent" in the last paragraph we mean that a given model function $f(t)=\sum A_jH_j(t)$ can be achieved with only one unique set of values for the pa-

rameters. If several different choices of the parameters all lead to the same model function, of course the data cannot distinguish between them; only certain functions of the parameters can be estimated accurately, however many data we have. In this case the parameters are said to be "confounded" or "unidentified". Generally, this would be a sign that the model was poorly chosen. However, it may be that the parameters are known to be real, and the experiment, whether by poor design or the perversity of nature, is just not capable of distinguishing them.

As an example of confounded parameters, suppose that two different sinusoidal signals are known to be present, but they have identical frequencies. Then their separate amplitudes are confounded: the data can give evidence only about their sum. The difference in amplitudes can be known only from prior information.

Chapter 4

ESTIMATING THE PARAMETERS

Once the models had been rewritten in terms of the orthonormal model functions, we were able to remove the nuisance parameters $\{A\}$ and σ. The integrals performed in removing the nuisance parameters were all Gaussian or gamma integrals; therefore, one can always compute the posterior moments of these parameters.

There are a number of reasons why these moments are of interest: the first moments of the amplitudes are needed if one intends to reconstruct the model $f(t)$; the second moments are related to the energy carried by the signal; the estimated noise variance σ^2 and the energy carried by the signal can be used to estimate the signal-to-noise ratio of the data. Thus the parameters $\{A\}$ and σ are not entirely "nuisance" parameters; it is of some interest to estimate them. Additionally, we cannot in general compute the expected value of the $\{\omega\}$ parameters analytically. We must devise a procedure for estimating these parameters and their accuracy.

4.1 The Expected Amplitudes $\langle A_j \rangle$

To begin we will compute the expected amplitudes $\langle A_j \rangle$ in the case where the variance is assumed known. Now the likelihood (3.12) is a function of the $\{\omega\}$ parameters and to estimate the $\langle A_j \rangle$ independently of the $\{\omega\}$'s, we should integrate over these parameters. Because we have not specified the model functions, we cannot do this once and for all. But we can obtain the estimate $\langle A_j \rangle$ as functions of the $\{\omega\}$ parameters. This gives us what would be the "best" estimate of the amplitudes if we

knew the $\{\omega\}$ parameters.

The expected amplitudes are given by

$$E(A_j|\{\omega\},\sigma,D,I) = \langle A_j\rangle = \frac{\int_{-\infty}^{+\infty} dA_1\cdots dA_m A_j L(\{A\},\{\omega\},\sigma)}{\int_{-\infty}^{+\infty} dA_1\cdots dA_m L(\{A\}|\{\omega\},\sigma)}.$$

We will carry out the first integration in detail to illustrate the procedure, and later just give results. Using the likelihood (3.12) and having no prior information about A_j, we assign a uniform prior, multiply by A_j and integrate over the $\{A\}$. Because the joint likelihood is a product of their independent likelihoods, all of the integrals except the one over A_j cancel:

$$\langle A_j\rangle = \frac{\int_{-\infty}^{+\infty} dA_j A_j \exp\left\{-\frac{1}{2\sigma^2}[A_j^2 - 2A_j h_j]\right\}}{\int_{-\infty}^{+\infty} dA_j \exp\left\{-\frac{1}{2\sigma^2}[A_j^2 - 2A_j h_j]\right\}}.$$

A simple change of variables $u_j = (A_j - h_j)/\sqrt{2\sigma^2}$ reduces the integrals to

$$\langle A_j\rangle = \frac{\int_{-\infty}^{+\infty} du_j\left\{\sqrt{2\sigma^2}u_j + h_j\right\}\exp\left\{-u_j^2\right\}}{\int_{-\infty}^{+\infty} du_j \exp\left\{-u_j^2\right\}}.$$

The first integral in the numerator is zero from symmetry and the second gives

$$\langle A_j\rangle = h_j. \tag{4.1}$$

This is the result one would expect. After all, we are expanding the data on an orthonormal set of vectors. The expansion coefficients are just the projections of the data onto the expansion vectors and that is what we find.

We can use these expected amplitudes $\langle A_j\rangle$ to calculate the expectation values of the amplitudes $\langle B_j\rangle$ in the nonorthogonal model. Using (3.10), these are given by

$$E(B_j|\{\omega\},\sigma,D,I) = \langle B_k\rangle = \sum_{j=1}^{m}\frac{h_j e_{jk}}{\sqrt{\lambda_j}}. \tag{4.2}$$

Care must be taken in using this formula, because the dependence of the $\langle B_k\rangle$ on the $\{\omega\}$ parameters is hidden. The functions h_j, the eigenvectors e_{kj} and the eigenvalues λ_j are all functions of the $\{\omega\}$ parameters. To remove the $\{\omega\}$ dependence from (4.2) one must multiply by $P(\{\omega\}|D,I)$ and integrate over all the $\{\omega\}$ parameters. If the total number of $\{\omega\}$ parameters r is large this will not be possible. Fortunately, if the total amount of data is large $P(\{\omega\}|D,I)$ will be so nearly a delta function that we can estimate these parameters from the maximum of $P(\{\omega\}|D,I)$.

Next we compute $\langle A_j \rangle$ when the noise variance σ^2 is unknown to see if obtaining independent information about σ will affect these results. To do this we need the likelihood $L(\{A\},\{\omega\})$; this is given by (3.12) after removing the variance σ^2 using a Jeffreys prior $1/\sigma$:

$$L(\{\omega\},\{A\}) \propto \left[\overline{d^2} - \frac{m\overline{h^2}}{N} + \frac{1}{N}\sum_{i=1}^{m}(A_j - h_j)^2\right]^{-\frac{N}{2}}. \tag{4.3}$$

Using (4.3) and repeating the calculation for $\langle A_j \rangle$ one obtains the same result (4.1). Apparently it does not matter if we know the variance or not. We will make the same estimate of the amplitudes regardless. As with some of the other results discovered in this calculation, this is what one's intuition might have said; knowing σ affects the accuracy of the estimates but not their actual values. Indeed, the first moments were independent of the value of σ when the variance was known; it is hard to see how the first moments could suddenly become different when the variance is unknown.

4.2 The Second Posterior Moments $\langle A_j A_k \rangle$

The second posterior moments $\langle A_j A_k \rangle$ cannot be independent of the noise variance σ^2, for that is what limits the accuracy of our estimates of the A_j. The second posterior moments, when the variance is assumed known, are given by

$$E(A_j A_k|\{\omega\},\sigma,D,I) = \langle A_j A_k \rangle = \frac{\int_{-\infty}^{+\infty} dA_1 \cdots dA_m A_j A_k L(\{A\},\{\omega\},\sigma)}{\int_{-\infty}^{+\infty} dA_1 \cdots dA_m L(\{A\},\{\omega\},\sigma)}.$$

Using the likelihood (3.12) and again assuming a uniform prior, these expectation values are given by

$$\langle A_j A_k \rangle = h_j h_k + \sigma^2 \delta_{jk}$$

so that, in view of (4.1), the posterior covariances are

$$\langle A_j A_k \rangle - \langle A_j \rangle \langle A_k \rangle = \sigma^2 \delta_{jk}. \tag{4.4}$$

The A_j parameters are uncorrelated (we defined the model functions $H_j(t)$ to ensure this), and each one is estimated to an accuracy $\pm\sigma$. Intuitively, we might anticipate this but we would not feel very sure of it.

The expectation value $\langle A_j A_k \rangle$ may be related to the expectation value for the original model amplitudes by using (3.11):

$$\langle B_k B_l \rangle - \langle B_k \rangle \langle B_l \rangle = \sigma^2 \sum_{j=1}^{m} \frac{e_{jk} e_{jl}}{\lambda_j}. \tag{4.5}$$

These are the explicit Bayesian estimates for the posterior covariances for the original model. These are the most conservative estimates (in the sense discussed before) one can make, but they are still functions of the $\{\omega\}$ parameters.

We can repeat these calculations for the second posterior moments in the case when σ is assumed unknown to see if obtaining explicit information about σ is of use. Of course, we expect the results to differ from the previous result since (4.5) depends explicitly on σ. Performing the required calculation gives

$$\begin{aligned} E(A_jA_k|\{\omega\},D,I) &= \langle A_jA_k\rangle = h_jh_k \\ &+ \left[\frac{N}{N-2}\right]\left[\frac{2N-5}{2N-5-2m}\right]\left[\frac{2N-7}{2N-7-2m}\right]\left[\overline{d^2}-\frac{m\overline{h^2}}{N}\right]\delta_{jk}. \end{aligned}$$

Comparing this with (4.4) shows that obtaining independent information about σ will affect the estimates of the second moments. But not by much, as we will see below. The second term in this equation is essentially an estimate of σ^2, but for small N it differs appreciably from the direct estimate found next.

4.3 The Estimated Noise Variance $\langle\sigma^2\rangle$

One of the things that is of interest in an experiment is to estimate the noise power σ^2. This indicates how "good" the data appear to be in the light of the model, and can help one in making many judgments, such as whether to try a new model or build a new apparatus. We can obtain the expected value of σ as a function of the $\{\omega\}$ parameters; however, we can just as easily obtain the posterior moments $\langle\sigma^s\rangle$ for any power s. Using (3.14), and the Jeffreys prior $1/\sigma$, we integrate:

$$E(\sigma^s|\{\omega\},D,I) = \langle\sigma^s\rangle = \frac{\int_0^{+\infty} d\sigma\sigma^{s-1}L(\sigma|\{\omega\},D,I)}{\int_0^{+\infty} d\sigma\sigma^{-1}L(\sigma|\{\omega\},D,I)}$$

to obtain

$$\langle\sigma^s\rangle = \Gamma\left(\frac{N-m-s}{2}\right)\Gamma\left(\frac{N-m}{2}\right)^{-1}\left[\frac{N\overline{d^2}-m\overline{h^2}}{2}\right]^{\frac{s}{2}}. \tag{4.6}$$

For $s=2$ this gives the estimated variance as

$$\begin{aligned} \langle\sigma^2\rangle &= \frac{N}{N-m-2}\left[\overline{d^2}-\frac{m\overline{h^2}}{N}\right] \\ &= \frac{1}{N-m-2}\left[\sum_{i=1}^{N} d_i^2 - \sum_{j=1}^{m} h_j^2\right]. \end{aligned} \tag{4.7}$$

The estimate depends on the number m of expansion functions used in the model. The more model functions we use, the smaller the last factor in (4.7), because by the Bessel inequality (3.18) the larger models fit the data better and $(\overline{d^2} - mN^{-1}\overline{h^2})$ decreases. But this should not decrease our estimate of σ^2 unless that factor decreases by more than we would expect from fitting the noise. The factor $N/(N-m-2)$ takes this into account. In effect probability theory tells us that $m+2$ degrees of freedom should go to estimating the model parameters and the variance, and the remaining degrees of freedom should go to the noise: everything not explicitly accounted for in the model is noise. We will show shortly that the estimated accuracy of the $\{\omega\}$ parameters depends directly on the estimated variance. If the model does not fit the data well, the estimates will become less precise in direct relation to the estimated variance.

We can use (4.6) to obtain an indication of the accuracy of the expected noise variance. The (mean) $\pm$ (standard deviation) estimate of σ^2 is

$$(\sigma^2)_{\text{est}} = \langle\sigma^2\rangle \pm \sqrt{\langle\sigma^4\rangle - \langle\sigma^2\rangle^2}.$$

From which we obtain

$$(\sigma^2)_{\text{est}} = \frac{N}{N-m-2}\left[\overline{d^2} - \frac{m\overline{h^2}}{N}\right]\left(1 \pm \epsilon\right)$$

$$\epsilon \equiv \sqrt{2/(N-m-4)}.$$

We then find the values of $N-m$ needed to achieve a given accuracy

% accuracy	ϵ	$N-m$
1	0.01	20,004
3	0.03	2,226
10	0.10	204
20	0.20	54

These are about what one would expect from simpler statistical estimation rules (the usual $N^{-\frac{1}{2}}$ rule of thumb).

4.4 The Signal-To-Noise Ratio

These results may be used to empirically estimate the signal-to-noise ratio of the data. We define this as the square root of the mean power carried by the signal

divided by the mean power carried by the noise:

$$\frac{\text{Signal}}{\text{Noise}} = \left[\langle \sum_{j=1}^{m} A_j^2 \rangle / N\sigma^2\right]^{\frac{1}{2}}.$$

This may be obtained from (4.2):

$$\frac{\text{Signal}}{\text{Noise}} = \left\{\frac{m}{N}\left[1 + \frac{\overline{h^2}}{\sigma^2}\right]\right\}^{\frac{1}{2}}. \tag{4.8}$$

A similar empirical signal-to-noise ratio may be obtained when the noise variance σ is unknown by replacing σ in (4.8) by the estimated noise variance (4.7). When the data fit the model so well that $\overline{h^2} \gg \sigma^2$, the estimate reduces to

$$\left\{\frac{m\overline{h^2}}{N\sigma^2}\right\}^{\frac{1}{2}} \quad \text{or} \quad \left\{\frac{\sum_{j=1}^{m} h_j^2}{\sum_{k=1}^{N} e_i^2}\right\}^{\frac{1}{2}}$$

We will compute the signal-to-noise ratio for several models in the following sections.

4.5 Estimating the $\{\omega\}$ Parameters

Unlike the amplitudes $\{A\}$ and the variance σ^2, we cannot calculate the expectation values of the $\{\omega\}$ parameters analytically. In general, the integrals represented by

$$\langle \omega_j \rangle = \int d\omega_1 \cdots d\omega_r \; \omega_j \; P(\{\omega\}|D, I)$$

cannot be done exactly. Nonetheless we must obtain an estimate of these parameters, and their probable accuracy.

The exact joint posterior density is (3.16) when σ is known, and (3.17) when it is not. But they are not very different provided *we have enough data for good estimates.* For, writing the maximum attainable $\sum h_j^2$ as

$$\left(\sum_{j=1}^{m} h_j^2\right)_{\max} = x$$

and writing the difference from the maximum as q^2 i.e.

$$\sum_{j=1}^{m} h_j^2 = x - q^2,$$

Eq. (3.17) becomes

$$\left[\sum_{i=1}^{N} d_i^2 - x + q^2\right]^{\frac{m-N}{2}} \approx \exp\left\{-\frac{(N-m)q^2}{2(\sum_{j=1}^{N} d_i^2 - x)}\right\}.$$

But this is nearly the same as

$$\left[\sum_{i=1}^{N} d_i^2 - x + q^2\right]^{\frac{m-N}{2}} \approx \exp\left\{-\frac{q^2}{2\langle\sigma^2\rangle}\right\}$$

where we used the estimate (4.7) for σ^2 evaluated for the values $\{\hat{\omega}\}$ that maximize the posterior probability as a function of the $\{\omega\}$ parameters. So up to an irrelevant normalization constant the posterior probability of the $\{\omega\}$ parameters around the location of the maximum is given by

$$P(\{\omega\}|D,\langle\sigma^2\rangle,I) \approx \exp\left\{\frac{m\overline{h^2}}{2\langle\sigma^2\rangle}\right\} \tag{4.9}$$

where the slightly inconsistent notation $P(\{\omega\}|\langle\sigma^2\rangle,D,I)$ has been adopted to remind us that we have used $\langle\sigma^2\rangle$, not σ^2. We have noted before that when we integrate over a nuisance parameter, the effect is for most purposes to estimate the parameter from the data, and then constrain the parameter to that value.

We expand $\overline{h^2}$, to obtain q^2, in a Taylor series around the maximum $\{\hat{\omega}\}$ to obtain

$$P(\{\omega\}|D,\langle\sigma^2\rangle,I) \propto \exp\left\{-\sum_{jk=1}^{r}\frac{b_{jk}}{2\langle\sigma^2\rangle}\Delta_j\Delta_k\right\} \tag{4.10}$$

where b_{jk} is the analogue of (2.9) defined in the single harmonic frequency problem

$$b_{jk} \equiv -\frac{m}{2}\frac{\partial^2\overline{h^2}}{\partial\omega_j\partial\omega_k} \tag{4.11}$$

$$\Delta_j \equiv \hat{\omega}_j - \omega_j.$$

From (4.10) we can make the (mean) $\pm$ (standard deviation) approximations for the $\{\omega\}$ parameters. We do these Gaussian integrals by first changing to orthogonal variables and then perform the r integrals just as we did with the amplitudes in Chapter 3. The new variables are obtained from the eigenvalues and eigenvectors of b_{jk}. Let u_{jk} denote the kth component of the jth eigenvector of b_{jk} and let v_j be the eigenvalue. The orthogonal variables are given by

$$s_j = \sqrt{v_j}\sum_{k=1}^{r}\Delta_k u_{kj} \qquad \Delta_j = \sum_{k=1}^{r}\frac{s_k u_{jk}}{\sqrt{v_k}}.$$

Making this change of variables, we have

$$P(\{s\}|\langle\sigma^2\rangle,D,I) \propto v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}}\exp\left\{-\sum_{j=1}^{r}\frac{s_j^2}{2\langle\sigma^2\rangle}\right\}. \tag{4.12}$$

From (4.12) we can compute $\langle s_j \rangle$ and $\langle s_j^2 \rangle$. Of course $\langle s_j \rangle$ is zero and the expectation value $\langle s_j s_k \rangle$ is given by

$$\langle s_k s_j \rangle = \frac{\int_{-\infty}^{\infty} ds_1 \cdots ds_r v_1^{-\frac{1}{2}} \cdots v_r^{-\frac{1}{2}} s_k s_j \exp\left\{-\sum_{l=1}^{r} \frac{s_l^2}{2\langle \sigma^2 \rangle}\right\}}{\int_{-\infty}^{\infty} ds_1 \cdots ds_r v_1^{-\frac{1}{2}} \cdots v_r^{-\frac{1}{2}} \exp\left\{-\sum_{l=1}^{r} \frac{s_l^2}{2\langle \sigma^2 \rangle}\right\}}$$

$$\langle s_k s_j \rangle = \langle \sigma^2 \rangle \delta_{kj}$$

where δ_{kj} is a Kronecker delta function. In the posterior distribution the s_j are uncorrelated, as they should be. From this we may obtain the posterior covariances of the $\{\omega\}$ parameters. These are

$$\langle \omega_j \omega_k \rangle - \langle \omega_j \rangle \langle \omega_k \rangle = \langle \sigma^2 \rangle \sum_{l=1}^{r} \frac{u_{lj} u_{lk}}{v_l},$$

and the variance γ_k^2 of the posterior distribution for ω_k is

$$\gamma_k^2 \equiv \langle \sigma^2 \rangle \sum_{j=1}^{r} \frac{u_{jk}^2}{v_j}. \tag{4.13}$$

Then the estimated ω_j parameters are

$$(\omega_j)_{\text{est}} = \hat{\omega}_j \pm \gamma_j \tag{4.14}$$

and; here $\hat{\omega}_j$ is the location of the maximum of the probability distribution as a function of the $\{\omega\}$ parameter.

For an arbitrary model the matrix b_{jk} cannot be calculated analytically; however, it can be evaluated numerically using the computer code given in Appendix E. We use a general searching routine to find the maximum of the probability distribution and then calculate this matrix numerically. The log of the "Student t-distribution" is so sharply peaked that gradient searching routines do not work well. We use a "pattern" search routine described by Hooke and Jeeves [21] [22].

The accuracy estimates of both the $\{\omega\}$ parameters and the amplitudes $\{A\}$ in Eq. (4.5) depend explicitly on the estimated noise variance. But the estimated variance is the mean square difference between the model and the data. If the misfit is large the variance is estimated to be large and the accuracy is estimated to be poor. Thus when we say that the parameter estimates are conservative we mean that, because everything probability theory cannot fit to the model is assigned to the noise, all of our parameter estimates are as wide as is consistent with the model and the

data. For example, when we estimate a frequency from a discrete Fourier transform we are in effect using a single harmonic frequency model for an estimate (position of a peak). But the width of the peak has nothing to do with the noise level, and if we supposed it, erroneously, to be an indication of the accuracy of our estimate, we could make very large errors.

This is perhaps one of the most subtle and important points about the use of uninformative priors that comes out in this work, and we will try to state it more clearly. When we did this calculation, at every point where we had to supply a prior probability we chose a prior that was as uninformative as possible (by uninformative we mean that the prior is as smooth as it can be and still be consistent with the known information). Specifically we mean a prior that has no sharp maximum: one that does not determine any value of the parameter strongly. We derived the Gaussian for the noise prior as the smoothest, least informative, prior that was consistent with the given second moment of the noise. We specifically did not assume the noise was nonwhite or correlated because we do not have prior information to that effect. So if the noise turns out to be colored we have in effect already allowed for that possibility because we used a less informative prior for the noise, which automatically considers every possible way of being colored, in the sense that the white noise basic support set includes all those of colored noise. On the other hand, if we knew a specific way in which the noise departs from whiteness, we could exploit that information to obtain a more concentrated noise probability distribution, leading to still better estimates of the $\{\omega\}$ parameters. We will demonstrate this point several times in Chapter 6.

4.6 The Power Spectral Density

Although not explicitly stated, we have calculated above an estimate of the total energy of the signal. The total energy E carried by the signal in our orthogonal model is

$$E \equiv \int_{t_1}^{t_N} f(t)^2 dt \approx \sum_{j=1}^{m} A_j^2$$

and its spectral density is given by

$$\hat{p}(\{\omega\}) = m\left[\sigma^2 + \overline{h^2}\right] P(\{\omega\}|D, I, \sigma). \tag{4.15}$$

This function is the energy per unit $\{\omega\}$ carried by the signal (not the noise). This power spectral estimate is essentially a power normalized probability distribution,

and should not be confused with what a power meter would measure (which is the total power carried by the signal and the noise).

We have seen this estimated variance term once before. When we derived the power spectral density for the single harmonic frequency a similar term was present [see Eq. (2.16)]. That term of $m\sigma^2$ in (4.15) might be a little disconcerting to some; if (4.15) estimates the energy carried by the "signal" why does it include the noise power σ^2? If $\overline{h^2} \gg \sigma^2$ then the term is of no importance. But in the unlikely event $\overline{h^2} \ll \sigma^2$, then what is this term telling us? When these equations were formulated we put in the fact that there is present noise of variance σ^2 in a space of dimension N, and a signal in a subspace of m model functions. But then if $\overline{h^2} \ll \sigma^2$, there is only one explanation: the noise is such that its components on those m model functions just happened to cancel the signal. But if the noise just cancelled the signal, then the power carried by the signal must have been equal to the power $m\sigma^2$ carried by the noise in those m functions; and that is exactly the answer one obtains. This is an excellent example of the sophisticated subtlety of Bayesian analysis, which automatically perceives things that our unaided intuition might not (and indeed did not) notice in years of thinking about such problems.

We have already approximated $P(\{\omega\}|D,\sigma,I)$ as a Gaussian expanded about the maximum of the probability density. Using (4.10) we can approximate the power spectral density as

$$\begin{aligned} \hat{p}(\{\omega\}) &\approx m[\langle\sigma^2\rangle + \overline{h^2}]P(\{\omega\}|\langle\sigma^2\rangle, D, I) \\ P(\{\omega\}|\langle\sigma^2\rangle, D, I) &\propto \exp\left\{-\sum_{jk=1}^{r} \frac{b_{jk}(\hat{\omega}_j - \omega_j)(\hat{\omega}_k - \omega_k)}{2\langle\sigma^2\rangle}\right\}. \end{aligned} \tag{4.16}$$

This approximation will turn out to be very useful. We will be dealing typically with problems where the $\{\omega\}$ parameters are well determined or where we wish to remove one or more of the $\{\omega\}$ parameters as nuisances. For example, when we plot the power spectral density for multiple harmonic frequencies, we do not wish to plot this as a function of multiple variables, but as a function of one frequency: all other frequencies must be removed by integration. We cannot do these integrals in (4.15); in general, however, we will be able to do them in (4.16).

There are two possible problems with this definition of the power spectral density. First we assumed there is only one maximum in the posterior probability density, and second we asked a question about the total power carried by the signal, not a question about one spectral line. It will turn out that the multiple frequency model will be invariant under permutations of the labels on the frequencies. It cannot matter

which frequency is number one and which is labeled number two. This invariance must manifest itself in the joint posterior probability density; there will be multiple peaks of equal probability and we will be led to generalize this definition. In addition we ask a question about the total energy carried by the signal in the sampling time. This is the proper question when the signal is not composed of sinusoids. But asking a question about the total energy is not the same as asking about the energy carried by each sinusoid. We will need to introduce another quantity that will describe the energy carried by one sinusoid. Before we do this we need to understand much more about the problem of estimating frequencies and decay rates. Chapter 6 is devoted primarily to this subject. For now we turn attention to a slightly more general problem of "how to make the optimal choice of a model?"

Chapter 5

MODEL SELECTION

When analyzing the results of an experiment it is not always known which model function (3.1) applies. We need a way to choose between several possible models. This is easily done using Bayes' theorem (1.3) and repeated applications of the procedure (1.4) which led to the "Student t-distribution." The first step in answering this question is to enumerate the possible models. Suppose we have a set of s possible models $\{H_1, \cdots, H_s\}$ with model functions $\{f_1, \cdots, f_s\}$. We are hardly ever sure that the "true" model is actually contained in this set. Indeed, the "set of all possible models" is not only infinite, but it is also quite undefined. It is not even clear what one could mean by a "true" model; both questions may take us into an area more like theology than science.

The only questions we seek to examine here are the ones that are answerable because they are mathematically well-posed. Such questions are of the form: "Given a specified set S_s of possible models $\{H_1, \cdots, H_s\}$ and looking only within that set, which model is most probable in view of all the data and prior information, and how strongly is it supported relative to the alternatives in that set?" Bayesian analysis can give a definite answer to such a question – see [15], [23].

5.1 What About "Something Else?"

To say that we confine ourselves to the set S_s is not to assert dogmatically that there are no other possibilities; we may assign prior probabilities $P(H_j|I), (1 \leq j \leq s)$

which do not add up to one:

$$\sum_{j=1}^{s} P(H_j|I) = a < 1.$$

Then we are assigning a prior probability $(1-a)$ to some unknown proposition

$$\text{SE} \equiv \text{“Something Else not yet thought of.”}$$

But until SE is specified it cannot enter into a Bayesian analysis; probability theory can only compare the specified models $\{f_1, \cdots, f_s\}$ with each other.

Let us demonstrate this more explicitly. If we try to include SE in our set of hypotheses, we can calculate the posterior probabilities of the $\{f_j\}$ and SE to obtain

$$P(f_j|D,I) = \frac{P(f_j|I)P(D|f_j,I)}{P(D|I)}$$

and

$$P(\text{SE}|D,I) = \frac{P(\text{SE}|I)P(D|\text{SE},I)}{P(D|I)}.$$

But this is numerically indeterminate even if $P(\text{SE}|I) = 1-a$ is known, because $P(D|\text{SE},I)$ is undefined until that “Something Else” is specified. The denominator $P(D|I)$ is also indeterminate, because

$$\begin{aligned} P(D|I) &= \sum_{j=1}^{s} P(D,f_j|I) + P(D,\text{SE}|I) \\ &= \sum_{j=1}^{s} P(D|f_j,I)P(f_j|I) + P(D|\text{SE},I)P(\text{SE}|I). \end{aligned}$$

But the relative probabilities of the specified models are still well defined, because the indeterminates cancel out:

$$\frac{P(f_i|D,I)}{P(f_j|D,I)} = \frac{P(f_i|I)}{P(f_j|I)}\frac{P(D|f_i,I)}{P(D|f_j,I)}.$$

These relative probabilities are independent of what probability $(1-a)$ we assign to “Something Else”, so we shall get the same results if we just ignore “Something Else” altogether, and act as if $a = 1$. In other words, while it is not wrong to introduce an unspecified “Something Else” into a probability calculation, no useful purpose is served by it, and we shall not do so here.

5.2 The Relative Probability of Model f_j

We wish to confine our attention to a selected set of models $\{f_1, \cdots, f_s\}$. Because of the arguments just given we may write

$$\sum_{j=1}^{s} P(f_j|D,I) = 1$$

where $P(f_j|D,I)$ is the posterior probability of model f_j. From Bayes' theorem (1.3) we may write

$$P(f_j|D,I) = \frac{P(f_j|I)P(D|f_j,I)}{P(D|I)} \tag{5.1}$$

and

$$P(D|I) = \sum_{j=1}^{s} P(f_j|I)P(D|f_j,I).$$

The way to proceed on this problem is to apply the general procedure for removing nuisance parameters given in Chapter 1. We will assume for now that the variance of the noise σ^2 is known and derive $P(f_j|\sigma,D,I)$, then at the end of the calculation if σ is not known we will remove it. Thus symbolically, we have

$$P(D|\sigma,f_j,I) = \int d\{A\}d\{\omega\}P(\{A\},\{\omega\}|I)P(D|\{A\},\{\omega\},\sigma,f_j,I). \tag{5.2}$$

But this is essentially just the problem we solved in Chapter 3 [Eqs. (3.12-3.17)] with three additions: when there can be differing numbers of parameters we must use normalized priors, we must do the integrals over the $\{\omega\}$ parameters, and the direct probability Eq. (5.2) of the data for the jth model must include all numerical factors in $P(D|\{A\},\{\omega\},\sigma,f_j,I)$. We will need to keep track of the normalization constants explicitly because the results we obtain will depend on them. We will do this calculation in four steps; first perform the integrals over the amplitudes $\{A\}$ using an appropriately normalized prior. Second we approximate the quasi-likelihood of the $\{\omega\}$ parameters about the maximum likelihood point; third remove the $\{\omega\}$ parameters by integration; and fourth remove the variances (plural because two more variances appear before we finish the calculation). Because the calculation is lengthy, we make many approximations of the kind that experienced users of applied mathematics learn to make. They could be avoided – but the calculation would then be much longer, with the same final conclusions.

We begin by the calculation in a manner similar to that done in Chapter 3. The question we would like to ask is "Given a set of model equations $\{f_1, \cdots, f_s\}$ and

looking only within that set, which model best accounts for the data?" We will take

$$f_j(t) = \sum_{k=1}^{m} A_k H_k(t, \{\omega\})$$

as our model, where H_k are the orthonormal model functions defined earlier, Eq. (3.5). The subscripts "j" refers to the jth member of the set of models $\{f_1, \cdots, f_s\}$, with the understanding that the amplitudes $\{A\}$, the nonlinear $\{\omega\}$, the total number of model functions m, and the model functions $H_k(t, \{\omega\})$ are different for every f_j. We could label each of these with an additional subscript, for example H_{jk} to stand for model function k of model f_j; however, we will not do this simply because the proliferation of subscripts would render the mathematics unreadable.

To calculate the direct probability of the data given model f_j we take the difference between the data and the model. This difference is the noise, if the model is true, and making the most conservative assumptions possible about the noise we assign a Gaussian prior for the noise. This gives the

$$P(D|\{A\}, \{\omega\}, \sigma, f_j, I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[d_i - f_j(t_i)]^2\right\}$$

as the direct probability of the data given model f_j and the parameters. Now expanding the square we obtain

$$P(D|\{A\}, \{\omega\}, \sigma, f_j, I) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{NQ}{2\sigma^2}\right\}$$

where

$$Q \equiv \overline{d^2} - \frac{2}{N}\sum_{l=1}^{m} A_l h_l + \frac{1}{N}\sum_{l=1}^{m} A_l^2$$

and

$$\sum_{i=1}^{N} H_l(t_i) H_k(t_i) = \delta_{lk}$$

and

$$h_l = \sum_{i=1}^{N} d_i H_l(t_i)$$

was used to simplify the expression. This is now substituted back into Eq. (5.2) to obtain

$$P(D|\sigma, f_j, I) = \int d\{A\} d\{\omega\} P(\{A\}, \{\omega\}|I)(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{NQ}{2\sigma^2}\right\}. \qquad (5.3)$$

At this point in the calculation we have simply repeated the steps done in Chapter 3 with one exception: we have retained the normalization constants in the direct

probability. To remove the amplitudes we assign an appropriate normalized prior and integrate. When we compare models with the same number of amplitudes and the same priors for them, the prior normalization factors do not matter: they simply cancel out of the posterior probability (5.1). But when we compare a model to one that has fewer amplitudes, these prior factors no longer cancel. We must keep track of them. In the calculation in Chapter 3 we used an improper uniform prior for these parameters. We cannot do that here because it smears out our prior information over an infinite range, and this would automatically exclude the larger model.

We will assume that the parameters are logically independent in the sense that gaining information about the amplitudes $\{A\}$ will not change our information about the nonlinear $\{\omega\}$ parameters, thus the prior factors:

$$P(\{A\},\{\omega\}|I) = P(\{A\}|I)P(\{\omega\}|I). \tag{5.4}$$

The amplitudes are location parameters and in Appendix A we derived an appropriate informative prior for a location parameter: the Gaussian. We will assume we have a vague previous measurement of the amplitudes $\{A\}$ and express this as a Gaussian centered at zero. Thus we take

$$P(\{A\}|\delta,I) = (2\pi\delta^2)^{-\frac{m}{2}} \exp\left\{-\sum_{k=1}^{m}\frac{A_k^2}{2\delta^2}\right\} \tag{5.5}$$

as our informative prior. In the Bayesian literature, δ is called a "hyperparameter". We will do this calculation for the case where we have little (effectively no) prior information: we assume $\delta^2 \gg \sigma^2$. That is, the prior measurement is much worse than the current measurement. Then the orthonormal amplitudes $\{A\}$ are all estimated with the same precision δ as required by Eq. (4.4).

Substituting the factored prior, Eq. (5.4), into Eq. (5.3) and then substituting the prior, Eq. (5.5), into Eq. (5.3) we arrive at

$$\begin{aligned} P(D|\delta,\sigma,f_j,I) &= \int d\{\omega\}P(\{\omega\}|I)(2\pi\delta^2)^{-\frac{m}{2}}(2\pi\sigma^2)^{-\frac{N}{2}} \\ &\times \int_{-\infty}^{+\infty} dA_1\cdots dA_m \exp\left\{-\sum_{k=1}^{m}\frac{A_k^2}{2\delta^2}\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\left[N\overline{d^2} - 2\sum_{k=1}^{m}A_k h_k + \sum_{k=1}^{m}A_k^2\right]\right\} \end{aligned}$$

as the direct probability of the data given model function f_j and the parameters. What is essential here is that the prior may be considered a constant over the range of values where the likelihood is large, but it goes to zero outside that range fast

enough to make it normalizable. Thus the last term in this integral looks like a delta function compared to the prior. We may write

$$\begin{aligned} P(D|\delta,\sigma,f_j,I) &= \int d\{\omega\}P(\{\omega\}|I)(2\pi\sigma^2)^{-\frac{N}{2}}(2\pi\delta^2)^{-\frac{m}{2}}\exp\left\{-\sum_{k=1}^{m}\frac{\hat{A}_k^2}{2\delta^2}\right\} \\ &\times \int_{-\infty}^{+\infty} d\{A\}\exp\left\{-\frac{1}{2\sigma^2}\Big[N\overline{d^2}-2\sum_{k=1}^{m}A_k h_k+\sum_{k=1}^{m}A_k^2\Big]\right\} \end{aligned}$$

where $\hat{A}$ is the location of the maximum of the likelihood as a function of the $\{A\}$ parameters. But from (4.1) $\hat{A}_j = h_j$ for a given model and after completing the integrals over the amplitudes, we have

$$\begin{aligned} P(D|\delta,\sigma,f_j,I) &= \int d\{\omega\}(2\pi\delta^2)^{-\frac{m}{2}}(2\pi\sigma^2)^{-\frac{(N-m)}{2}}P(\{\omega\}|I) \\ &\times \exp\left\{-\frac{N\overline{d^2}-m\overline{h^2}}{2\sigma^2}-\frac{m\overline{h^2}}{2\delta^2}\right\} \end{aligned}$$

as the direct probability of the data given the model function f_j and the parameters.

The second step in this calculation is to approximate $\overline{h^2}$ around the maximum and then perform the integrals over the $\{\omega\}$ parameters. The prior uncertainty $\delta \gg \sigma$, so the prior factor in the above equation is only a small correction. When we expand $\overline{h^2}$ about the maximum $\{\hat{\omega}\}$ we will not bother expanding this term. This permits us to use the same approximation given earlier (4.10, 4.9) while making only a small error. We Taylor expand $\overline{h^2}$ to obtain

$$\begin{aligned} P(D|\delta,\sigma,f_j,I) &\approx \int d\{\omega\}P(\{\omega\}|I)(2\pi\delta^2)^{-\frac{m}{2}}(2\pi\sigma^2)^{-\frac{N-m}{2}} \\ &\times \exp\left\{-\frac{N\overline{d^2}-m\overline{h^2}(\{\hat{\omega}\})}{2\sigma^2}-\frac{m\overline{h^2}(\{\hat{\omega}\})}{2\delta^2}\right\} \\ &\times \exp\left\{-\sum_{k,l}^{r}\frac{b_{kl}(\hat{\omega}_k-\omega_k)(\hat{\omega}_l-\omega_l)}{2\sigma^2}\right\}. \end{aligned} \tag{5.6}$$

We are now in a position to remove the $\{\omega\}$ parameters. To do this third step in the calculation we will again assign a normalized prior for them. When we Taylor expanded $\overline{h^2}$ we made a local approximation to the direct probability of the data given the parameters. In this approximation the $\{\omega\}$ parameters are location parameters. We again assume a prior which is Gaussian with some variance γ, another hyperparameter. We have

$$P(\{\omega\}|\gamma,I) = (2\pi\gamma^2)^{-\frac{r}{2}}\exp\left\{-\sum_{k=1}^{r}\frac{\omega_k^2}{2\gamma^2}\right\} \tag{5.7}$$

as the informative prior for the $\{\omega\}$ parameters. If the $\{\omega\}$ parameters are frequencies then one could argue that they are scale parameters, for which the completely uninformative prior is the nonnormalizable Jeffreys prior; and so we should choose a normalizable prior that resembles it. However, that does not matter; the only properties of our prior that survive are the prior density at the maximum likelihood point and the prior range, and even these may cancel out in the end. We are simply playing it safe by using normalized priors so that no singular mathematics can arise in our calculation; and it does not matter which particular ones we use as long as they are broad and uninformative.

Substituting the prior (5.7) into Eq. (5.6), the integral we must perform becomes

$$\begin{aligned} P(D|\gamma,\delta,\sigma,f_j,I) &\approx \int d\{\omega\}(2\pi\delta^2)^{-\frac{m}{2}}(2\pi\gamma^2)^{-\frac{r}{2}}(2\pi\sigma^2)^{-\frac{N-m}{2}} \\ &\times \exp\left\{-\frac{N\overline{d^2}-m\overline{h^2}(\{\hat{\omega}\})}{2\sigma^2}-\frac{m\overline{h^2}(\{\hat{\omega}\})}{2\delta^2}-\sum_{k=1}^{r}\frac{\omega_k^2}{2\gamma^2}\right\} \\ &\times \exp\left\{-\sum_{k,l}^{r}\frac{b_{kl}(\hat{\omega}_k-\omega_k)(\hat{\omega}_l-\omega_l)}{2\sigma^2}\right\}. \end{aligned}$$

We will again assume that the prior information is vague, $\gamma \gg \sigma$, we treat the last term in the integral like a delta function compared to the prior. Thus we will take the prior factors out of the integral and simply evaluate them at the maximum likelihood point. Then integrating over the $\{\omega\}$ parameters gives the direct probability of the data given the model f_j and the three remaining parameters. If these three parameters are actually known then the direct probability is given by

$$\begin{aligned} P(D|\gamma,\delta,\sigma,f_j,I) &= (2\pi\delta^2)^{-\frac{m}{2}}\exp\left\{-\frac{m\overline{h^2}(\{\hat{\omega}\})}{2\delta^2}\right\} \\ &\times (2\pi\gamma^2)^{-\frac{r}{2}}\exp\left\{-\frac{r\overline{\omega^2}}{2\gamma^2}\right\}v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}} \\ &\times (2\pi\sigma^2)^{-\frac{N-m-r}{2}}\exp\left\{-\frac{N\overline{d^2}-m\overline{h^2}(\{\hat{\omega}\})}{2\sigma^2}\right\} \end{aligned} \tag{5.8}$$

where $\overline{\omega^2} = (1/r)\sum_{k=1}^{r}\hat{\omega}^2$ is the mean-square $\{\hat{\omega}\}$ for model f_j, and $\overline{h^2}(\{\hat{\omega}\})$ is the mean-square projection of the data onto the orthonormal model functions evaluated at the maximum likelihood point for model f_j, and $v_1^{-\frac{1}{2}}\cdots v_r^{-\frac{1}{2}}$ is the Jacobian introduced in Eq. (4.12). If the three variances are known then the problem is complete, and the number which must be used in (5.1) is given by (5.8).

We noted earlier that one must be careful with the prior factors when the models have different numbers of parameters and we can see that here. If two models have

different values of m or r, their relative likelihood will have factors of the form $(2\pi\delta^2)^x$, or $(2\pi\gamma^2)^y$. The prior ranges remain relevant, a fact that we would have missed had we used improper priors.

There are three variances, σ, δ, and γ, in the direct probability of the data. We would like to remove these from $P(D|\gamma,\delta,\sigma,f_j,I)$. We could remove these using a Jeffreys prior, because each of these parameters appears in every model. The infinity introduced in doing this would always cancel out formally. However, to be safe, we can bound the integral, normalize the Jeffreys prior, and then remove these variances; then even if the normalization constant did not cancel we would still obtain the correct result. We will proceed with this last approach. There are three variances, and therefore three integrals to perform. Each of the three integrals is of the form:

$$\frac{1}{\log(H/L)}\int_L^H ds\frac{s^{-a}\exp\left\{-\frac{Q}{s^2}\right\}}{s}$$

where H stands for the upper bound on the variance, L for the lower bound, $\log(H/L)$ is the normalization constant for the Jeffreys prior, s is any one of the three variances, and Q and a are constants associated with s. A change of variables $u = Q/s^2$ reduces this integral to

$$\frac{1}{2}\frac{Q^{-\frac{a}{2}}}{\log(H/L)}\int_{\sqrt{\frac{Q}{H}}}^{\sqrt{\frac{Q}{L}}} du\ u^{\frac{a}{2}-1}\mathrm{e}^{-u}.$$

This integral is of the form of an incomplete Gamma integral. But our knowledge of the limits on this integral is vague: we know only that L is small and H large. If, for example, we assume that

$$\sqrt{\frac{Q}{H}} \ll 1 \quad \text{and} \quad \frac{a}{2}-1 \ll \sqrt{\frac{Q}{L}}$$

then the integrand is effectively zero at the limits; we can take the integral to be approximately $\Gamma(a/2)$. Designating the ratio of the limits H/L as R_α, where the subscript represents the limits for σ, δ, or γ integral, the three integrals are given approximately by

$$\int_L^H d\delta\frac{\delta^{-m}\exp\left\{-m\overline{h^2}/2\delta^2\right\}}{\log(R_\delta)\delta} \approx \frac{\Gamma(m/2)}{2\log(R_\alpha)}\left[\frac{m\overline{h^2}}{2}\right]^{-\frac{m}{2}}$$

for δ and

$$\int_L^H d\gamma\frac{\gamma^{-r}\exp\left\{-r\overline{\omega^2}/2\gamma^2\right\}}{\log(R_\gamma)\gamma} \approx \frac{\Gamma(r/2)}{2\log(R_\gamma)}\left[\frac{r\overline{\omega^2}}{2}\right]^{-\frac{r}{2}}$$

for γ and

$$\int_L^H d\sigma \frac{\sigma^{m+r-N} \exp\left\{-[N\overline{d^2} - m\overline{h^2}]/2\sigma^2\right\}}{\log(R_\sigma)\sigma} \approx \frac{\Gamma(\frac{N-m-r}{2})}{2\log(R_\sigma)} \left[\frac{N\overline{d^2} - m\overline{h^2}(\{\omega\})}{2}\right]^{\frac{m+r-N}{2}}$$

for σ.

Using these three integrals the global likelihood of the data given the model f_j is given by

$$\begin{aligned} P(D|f_j, I) &= \frac{\Gamma(m/2)}{2\log(R_\delta)} \left[\frac{m\overline{h^2}(\{\hat{\omega}\})}{2}\right]^{-\frac{m}{2}} \frac{\Gamma(r/2)}{2\log(R_\gamma)} \left[\frac{r\overline{\omega^2}}{2}\right]^{-\frac{r}{2}} v_1^{-\frac{1}{2}} \cdots v_r^{-\frac{1}{2}} \\ &\times \frac{\Gamma([N-m-r]/2)}{2\log(R_\sigma)} \left[\frac{N\overline{d^2} - m\overline{h^2}(\{\omega\})}{2}\right]^{\frac{m+r-N}{2}}. \end{aligned} \tag{5.9}$$

The three factors involved in normalizing the Jeffreys priors appear in every model: they always cancel as long as we deal with models having all three types of parameters. However, as soon as we try to compare a model involving two types of parameters to a model involving all three types of parameters (e.g. a regression model to a nonlinear model) they no longer cancel: the prior ranges become important. One must think carefully about just what prior information one actually has about γ, and δ, and use that information to set their prior ranges. As we shall see in what follows, if the data actually determine the model parameters well (so that these equations apply) the actual values one assigns to δ and γ are relatively unimportant.

5.3 One More Parameter

We would like to understand (5.1), (5.8), and (5.9) better, and we present here a simple example of their use. Suppose we are dealing with the simplest model selection possible: expanding the data on a set of model functions. A typical set of model functions might be polynomials. This is the simplest model possible because there are no $\{\omega\}$ parameters; only amplitudes. But suppose further that we choose our model functions so that they are already orthogonal in the sense defined earlier. All that is left for us to decide is "When have we incorporated enough expansion vectors to adequately represent the signal?" We will assume in this demonstration that both the variance σ and the prior variance δ are known and apply (5.8). Further, we will be comparing only two models at a time, and will compute the ratio of Eq. (5.8) for a model containing m expansion functions (or vectors on the discrete sample points)

to a model containing $m+1$ expansion functions. This ratio is called the posterior "odds" in favor of the smaller model.

When we have m expansion vectors and no $\{\omega\}$, Eq. (5.8) reduces to

$$\begin{aligned} P(D|f_m,\sigma,\delta,I) &= (2\pi\delta^2)^{-\frac{m}{2}}\exp\left\{-\sum_{k=1}^{m}\frac{h_k^2}{2\delta^2}\right\} \\ &\times (2\pi\sigma^2)^{-\frac{N-m}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[N\overline{d^2}-\sum_{k=1}^{m}h_k^2\right]\right\} \end{aligned}$$

for the first model and to

$$\begin{aligned} P(D|f_{m+1},I) &= (2\pi\delta^2)^{-\frac{m+1}{2}}\exp\left\{-\sum_{k=1}^{m+1}\frac{h_k^2}{2\delta^2}\right\} \\ &\times (2\pi\sigma^2)^{-\frac{N-m-1}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[N\overline{d^2}-\sum_{k=1}^{m+1}h_k^2\right]\right\} \end{aligned}$$

for a model with $m+1$ parameters. Because these models are already orthonormal, h_k is the same in both equations: when we compute the odds ratio all but the last will cancel. Thus the posterior odds ratio simplifies considerably. We have the likelihood ratio

$$L=\frac{P(D|f_m,\sigma,\delta,I)}{P(D|f_{m+1},\sigma,\delta,I)}=\frac{\delta}{\sigma}\exp\left\{\left(\frac{1}{\delta^2}-\frac{1}{\sigma^2}\right)\frac{h_{m+1}^2}{2}\right\}.$$

The posterior odds ratio then involves the posterior probabilities:

$$\frac{P(f_m|\sigma,\delta,D,I)}{P(f_{m+1}|\sigma,\delta,D,I)}=\frac{P(f_m|I)}{P(f_{m+1}|I)}L.$$

We derived this approximation assuming $\delta\gg\sigma$, so we have

$$L=\frac{\delta}{\sigma}\exp\left\{-\frac{h_{m+1}^2}{2\sigma^2}\right\}.$$

In other words, the smaller model is helped by uncertainty in the prior knowledge of A_{m+1}, while the larger model is helped by the relative size of the estimated next amplitude compared to the noise. This is the Bayesian quantitative version of Occam's razor: prefer the simpler model unless the bigger one achieves a significantly better fit to the data. For the bigger model to be preferred, the $m+1$ model function's projection onto the data must be large compared to the noise. Thus the Bayesian answer to this question essentially tells one to do what his common sense might have told him to do. That is, to continue increasing the number of expansion vectors until the projection of the data onto the next vector becomes comparable to the noise.

But we can be more specific than this. For example assume that $100\sigma = \delta$. Then to achieve $L = 1$, we need

$$\begin{aligned} \log(100) - \frac{h_{m+1}^2}{2\sigma^2} &= 0 \\ h_{m+1} &= \pm 3.0\sigma. \end{aligned}$$

The "data fitting factor" cancels out the "Occam factor" when the next projection is three times the RMS noise. Projections larger than this will favor the more complicated model.

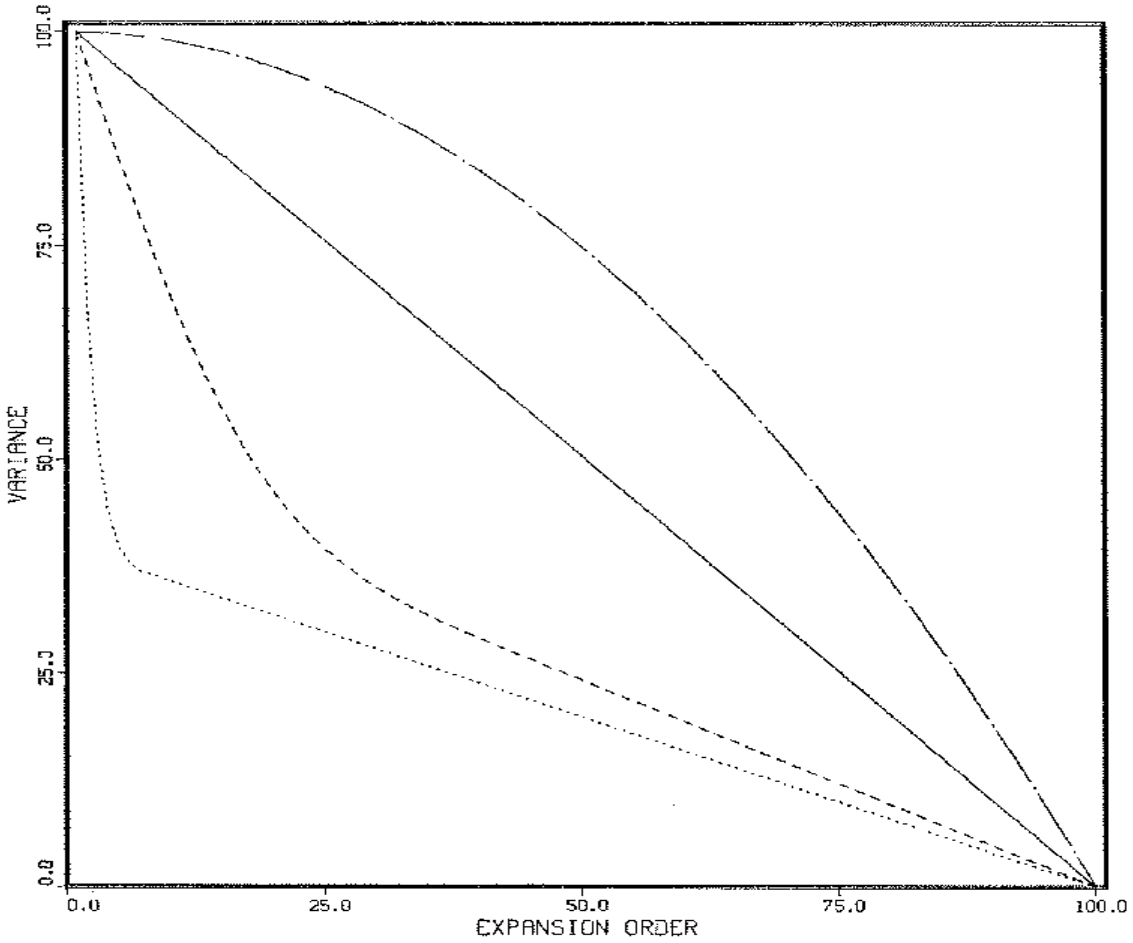
This result does not depend strongly on the assumed prior information. Here we took δ to be 100 times larger than σ. But the answer depends on the square root of $\log(\delta)$. So even if δ had been a billion (10^9) times larger it would have increased the critical value of h_{m+1} only by a factor of 2.3. Thus probability theory can give one a reasonable criterion for choosing between models, that depends only weakly on the prior information. There is hardly any real problem in which one would not feel sure in advance that $\delta < 10^{11}\sigma$, and few in which that 10^{11} could not be reduced to 10^2. But to try to go an improper prior $\delta \to \infty$, would give entirely misleading results; the larger model could never be accepted, whatever the data. Thus, while use of proper priors is irrelevant in many problems, it is mandatory in some.

5.4 What is a Good Model?

We can now state what we mean by a good model. We know from the Bessel inequality (3.17) that the estimated noise variance will have a value of $\overline{d^2}$ when we have no model functions. As we include more model functions, the estimated variance must go monotonically to zero. We can plot the estimated variance as a function of the expansion order, Fig. 5.1 (by expansion order we mean the total number of model functions m).

There are three general regions of interest: First, the solid line running from $\overline{d^2}$ down to zero (we will call this a "bad" model); second the region with values of σ^2 below this line; and third the region above this line. The region above the line is not a bad or a good region; it is simply one in which the model functions have been labeled in a bad order. By reordering the model functions we will obtain a curve below the straight line.

Figure 5.1: Choosing a Model



The solid line represents the worst possible choice of model functions. The region above this line is neither good nor bad (see text). The region below the line represents the behavior of good models. One strives to obtain the largest drop in the estimated variance with the fewest model functions. The dashed line might represent a fair model and the dotted line the "best" model.

Let $\Delta(\langle\sigma^2\rangle)$ stand for the change in the estimated variance σ^2 from incorporating one additional model function [we define $\Delta(\langle\sigma^2\rangle)$ to be positive]. We assume the model functions are incorporated in order of decreasing $\Delta(\langle\sigma^2\rangle)$: the model function with the largest $\Delta(\langle\sigma^2\rangle)$ is labeled one; the model function which produces the second largest $\Delta(\langle\sigma^2\rangle)$ is number two, etc.

We called the solid line a "bad" model because all of the $\Delta(\langle\sigma^2\rangle)$'s are the same; there is no particular model function which resembles the data better than any other. It would require outstandingly bad judgment – or bad luck – to choose such a set of model functions. But something like the linear behavior is to be expected when one expands pure noise on a complete set. On the other hand, if there is a signal present one expects to do better than this until the signal has been expanded; then one expects the curve to become slowly varying.

We can characterize a model by how quickly the $\Delta(\langle\sigma^2\rangle)$ drops. Any curve which drops below another curve indicates a model which is better, in the sense that it achieves a better quality of fit to the data with a given number of model functions. The "best" model is one which projects as much mean-square data as possible onto the first few model functions. What one would expect to find is: a very rapid drop as the systematic signal is taken up by the model, followed by a slow drop as additional model functions expand the noise.

We now have the following intuitive picture of the model fitting process: one strives to find models which produce the largest and fastest drop in $\langle\sigma^2\rangle$; any model which absorbs the systematic part of the signal faster than another model is a better model; the "best" is one which absorbs all of the systematic part of the signal with the fewest model parameters. This corresponds to the usual course of a scientific research project; initially one is very unsure of the phenomenon and so allows many conceivable unknown parameters with a complicated model. With experience one learns which parameters are irrelevant and removes them, giving a simpler model that accounts for the facts with fewer model functions. The total number of "useful" model functions is determined by the location of the break in the curve. The probability of any particular model can be computed using (4.15), and this can be used to estimate where the break in the curve occurs.

Of course, in a very complicated problem, where the data are contaminated by many spurious features that one could hardly hope to capture in a model, there may not be any well-defined breaking point. Even so, the curve is useful in that its shape gives some indication of how clean-cut the problem is.