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

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

New York Berlin Heidelberg London Paris Tokyo

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 Sunchez, 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

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

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

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

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

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

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

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.