Bayesian inversion with applications to physics

Bayesian inversion with applications to physics

Vincent A. Macaulay

Merton College
Trinity Term 1992

A thesis submitted in partial fulfillment
of the requirements for the
degree of Doctor of Philosophy

Theoretical Physics
University of Oxford

(Submitted 31 July 1992)

For my parents

When I began to muse about probability,
everything first seemed to be so simple and clear,
and only now do I realize how wrong I was!
Every time I believe I have grasped the truth,
it again slips out of my hands.
There are lurking abysses at almost every step of the way.
Alfréd Rényi: Levelek a valószínűségről (Letters on Probability)

'The time has come,' the Walrus said,
'To talk of many things.'
Lewis Carroll: Through the Looking-Glass and What Alice Found There

Acknowledgements

I have great pleasure in thanking Brian Buck, my supervisor over the past four years and the person who initiated me into the Bayesian Mysteries when I was still an undergraduate in Oxford. I hope he has enjoyed the many discussions we have had as much as I have; I have learnt a lot. He kindly read through this thesis and made many useful suggestions.

Helpful discussions with Geoff Daniell, Steve Gull, David MacKay, Brett Patterson, Prasenjit Saha and John Skilling are acknowledged.

Funding from the Science and Engineering Research Council, Merton College, the Scott Fund and the Subdepartment of Theoretical Physics made this work possible, not least by allowing me to attend three MaxEnt conferences abroad.

Thanks are due to Paul Aspinwall for so effectively looking after the departmental workstations, on which this thesis rather heavily relied, both for the computations and the typesetting.

Finally, to the dedicatees, I send my deepest love and thanks for their support. I hope this thesis goes some way to explaining what I have been doing for the last few years.

V.A.M.
Oxford, July 1992
`vincent@dionysos.thphys.ox.ac.uk`

Bayesian inversion with applications to physics
Vincent A. Macaulay
Merton College

Submitted for the degree of Doctor of Philosophy
Trinity Term 1992

Abstract

The estimation of a density function from noisy measurements of some integral transform linear in that density is a common problem in physics. After discussing the nature of inference and its natural algebra, probability theory, I show how to apply it to this type of inverse problem. A crucial step in formulating the solution is the identification of a suitable prior probability for density functions, namely, the entropic form.

The problem of the reconstruction of nuclear charge densities from scattering data is tackled. In order to do this, a new prior of the entropic family is proposed for the inference of distributions which are not known to be positive *a priori*. The reliability of the inferred reconstruction and of any integral linear in the density is determined by calculating error-bars. It is shown that the computational expense of this calculation can be reduced from an image-sized problem to one governed by the number of data points. I discuss the interpretation of error-bars and demonstrate that, quite reasonably, they depend strongly on the fineness of the detail that one is trying to infer. The results show conclusively that it is possible to infer charge densities without having to concoct explicit models of their shape or to make assumptions about their smoothness, which has been the case in the past. The theory has countless potential applications.

A technique for inferring the spectral density of noisy time-series is developed. It is applied to the famous Wolf sunspot index and its increased resolving power, compared to other methods, combined with a plausible preprocessing of the data, has allowed the identification of new features in this record. As well as a fundamental multiplet, a time-delayed third harmonic is seen, which can be explained by a cubic function acting on the whole fundamental. A second group of spectral features shadows the fundamental with frequencies increased approximately 1.6 times. This curious factor is suggestive of non-linear processes occurring in the sun.

Contents

Introduction

This thesis tries to demonstrate the power of using Bayes' theorem to infer density distributions with a prior probability assigned from the entropic family. In the past few years, this theory, known as MaxEnt, has been brought into such a form that it is now easy to see how to use it in problems which have their own idiosyncrasies and which, as a result, require changes to individual elements of that theory, such as the prior.

The first chapter is a crash course in Bayesian probability theory, in the interpretation favoured by Jeffreys, Cox, Jaynes, Skilling and Gull, among many others. It attempts briefly to introduce all the techniques and vocabulary that are a necessary background to the rest of the work. I describe the rôle that the theory performs, namely, a general algebra of plausible reasoning, and introduce the reader to the sum and product rules, Bayes' theorem and marginalization. The assignment of probabilities is touched on, and the part that the entropy function has to play in that assignment discussed.

In Chapter 2, I show that the reconstruction of a function, the data on which consists of integrals linear in that function, is achieved using Bayes' theorem, just as is any other problem of inferring quantities from experimental measurements. The tricky part is to identify a useful prior distribution to place on the function. That the prior is a necessity is demonstrated by considering carefully the limitations of the likelihood in this kind of inverse problem, where there are many more variables to infer than there are data points. A plausible argument is given to show that an entropic form is a reasonable choice for the prior and it is demonstrated that, in most situations, it betters the Gaussian, the predictions of which are worked out in detail. Although it is not possible to solve the entropic equations for the most probable image exactly, insight into their solution is obtained by showing how the posterior distribution extrapolates from the given data into regions where no data was measured. I demonstrate how these equations can be projected into a space of a size governed only by the number of data points, and how the best image can be obtained from that space by exponentiation. I briefly discuss how the regularization constant can be determined and also how the best image is computed.

Chapter 3 describes the reconstruction of nuclear charge densities from elastic electron scattering data using the methods described in Chapter 2. A new entropic prior is derived which is suitable for functions which can be positive or negative and hence does not rule out regions of negative charge density in the nucleus. Most previous inversions of this kind of data have been accomplished by fitting specific models of the nuclear shape to the data. The models come either from theories about the nucleus or more commonly are merely plausible parametrizations of the nuclear shape. The reconstructions stand or fall by the truth of the model employed. Thus it is useful to try to reconstruct the density making as few assumptions about the nucleus as possible. I develop the machinery to calculate error-bars around the most probable reconstruction and discuss how they are to be interpreted. I show that in the continuum limit, as the discretized image turns into a continuous function and error-bars into error-bands, the errors diverge. This is interpreted as indicating that structure on the finest scale is not determined by the data, which is a common sense result. The theory is demonstrated on the two light nuclei He^4 and C^{12}.

In the chapter on spectral estimation, a technique is developed for reconstructing spectral density functions from noisy time-series. It is a tool which has been missing from the Bayesian spectral estimation toolkit for some time. Solution to two different problems in this family are already available. The estimation of the spectral density of a signal from reliable noiseless measurements of its fluctuating autocorrelation function is done using Burg's method, while parameter estimation and comparison of models of a signal containing a small number of sinusoidal components has been solved by Jaynes and Bretthorst. A solution to the task of inferring the most probable spectrum when the data are noisy and incomplete is presented here. It is an inverse problem of the type that was discussed in Chapter 2, and is solved, as always, by assigning a prior and likelihood to the spectrum. I explain why the prior can only be a function of the amplitude at a particular frequency and not of the phase. This prior is taken from the entropic family. The posterior is the probability of a particular spectrum and allows the inference of amplitudes and phases. It is a Bayesian replacement for such orthodox estimators as the discrete Fourier transform or the Schuster periodogram. As expected, it displays the usual entropic benefits of noise and artifact suppression.

In Chapter 5, the spectral techniques developed in the previous chapter are used in an investigation of the Wolf sunspot index. This is a measure of the sun's activity, for which nearly 300 years of data is available. The case is made that this data set in fact represents a rectified version of the driving 'field', and it is confirmed that derectification of the data simplifies the interpretation of the signal remarkably. The improved resolution and cleanness of the entropically restored spectrum allow a number of new features of the data to be identified. The most important periodicity is the well-known 22.1 year cycle. But this spectral line is only one of at least four closely equally-spaced lines in the fundamental region. A third harmonic can be identified and a model of it made by taking a cubic delayed function of the fundamental; it is manifest in the time-series as regular jagged features on the trailing edges of cycles. Most suggestive of all is a group of lines which mimic the fundamental multiplet but at frequencies increased approximately 1.6 times. I speculate that this can only have an explanation in a non-linear theory of the solar activity cycle. In addition, the spectral signature of a curious anomaly which occurred in the data between 1775 and 1810 is discussed and I suggest that it might be the sign of a pulse repeated every 200 years which drives the solar cycle.

1
Reasoning under uncertainty

"Why, you might just as well say that 'I see what I eat'
is the same thing as 'I eat what I see'!"
"You might just as well say that 'I like what I get'
is the same thing as 'I get what I like'!"
"You might just as well say that 'I breathe when I sleep'
is the same thing as 'I sleep when I breathe'!"
Lewis Carroll: Alice's Adventures in Wonderland

This chapter will review the Bayesian probability theory, the only consistent mathematical apparatus we have for making rational inferences when we do not have complete information. This theory underpins all of the work in this thesis and the necessary parts of it are discussed here. These include Bayes' theorem, the marginalization rule, and the assignment of probabilities by maximizing entropy. Special probabilities such as the prior and the likelihood are defined.

1.1 Inference in science

As physical scientists, we desire to make propositions about the world we live in, which we believe can be true or false. Experiments are conducted, perhaps by a particle physicist at a collider facility or by a theorist in the memory of a digital computer, to try to determine which is the more likely; rarely, if ever, can a series of logical deductions be made, leading from data to theory. For, uncertainties creep into experiments, however carefully they are performed and finite resources limit how much data can be taken. And yet conclusions *are* made as a result of experiments; good theories *are* separated from bad. For instance, we feel progressively more confident in the theory of General Relativity as its predictions compare well with experiment decade by decade, although no experiment has proved or ever will prove it.

The method by which we reason in these circumstances of incomplete information is called **inference**[1] and a unique mathematical language which

[1] Polya (1954) gives a good account of the general properties a theory of inference should have.

describes its essential features is available in the Bayesian probability theory. This theory arises out of very simple considerations of what is meant by the likeliness[2] or plausibility of a proposition. The idea is to set up a scale on which these plausibilities can be ranked for a person in some given state of knowledge and to show how these numbers can be assigned, and modified in the light of new evidence.

Propositions will be denoted by $\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$. Their meaning should be unambiguous and they should be precise enough that it is possible for them to be true or false. For example, a valid $\mathcal{A}$ might be *"The gravitational constant has a value between* 6.67×10^{-11} *and* $6.68 \times 10^{-11}\,\mathrm{N\,m^2\,kg^{-2}}$*"*, which we believe is very likely true. But *"Wilkie Collins was a better novelist than Dickens"* is much more problematical for our purposes because 'better' is likely to mean different things to different people (Jaynes, in preparation).

From a single proposition, $\mathcal{A}$ say, its negation $\overline{\mathcal{A}}$ can be formed, which asserts the opposite of $\mathcal{A}$. If $\mathcal{A}$ is true, then $\overline{\mathcal{A}}$ is false and *vice versa.* In addition, from a set of propositions, we can make many new ones by combining them in various ways. The *conjunction* of $\mathcal{A}$ and $\mathcal{B}$ is written $\mathcal{A} \cdot \mathcal{B}$ and is the assertion of both $\mathcal{A}$ and $\mathcal{B}$, while the *disjunction* of a pair of propositions is written $\mathcal{A} + \mathcal{B}$ and asserts one or other or both.

One thing should be apparent straightaway and that is that the plausibility of any proposition depends not just on the proposition itself but on the truth values of a host of others, which are the information that allows us to judge its plausibility in the first place. For example, I might assign a very different plausibility to the proposition that *"Tomorrow the* FTSE-100 *share index will fall by more than 50 points"* based on propositions which state that today was either a very good or a very bad day on Wall Street. All plausibilities are necessarily conditional objects. We acknowledge this by writing the plausibility of $\mathcal{A}$ based on the truth of the proposition $\mathcal{C}$ as $p(\mathcal{A}|\mathcal{C})$ and speak of the plausibility of $\mathcal{A}$ conditioned on $\mathcal{C}$.

1.2 The algebra of probable inference

1.2.1 Cox's axioms

It was Cox (1946, 1961) who showed that, if a calculus of the plausibilities of propositions was to exist, the plausibilities had to obey the sum and product rules of probabilities,[3] respectively,

$$\mathrm{prob}(\mathcal{A}|\mathcal{C}) + \mathrm{prob}(\overline{\mathcal{A}}|\mathcal{C}) = 1 \qquad \text{and} \tag{1.1}$$

$$\mathrm{prob}(\mathcal{A} \cdot \mathcal{B}|\mathcal{C}) = \mathrm{prob}(\mathcal{A}|\mathcal{C})\,\mathrm{prob}(\mathcal{B}|\mathcal{A} \cdot \mathcal{C}). \tag{1.2}$$

[2] Not to be confused with *likelihood*, which is a special sort of probability.

[3] In fact they can obey slightly more general equations, but it is a generality which adds nothing to the theory—see Cox (1961).

In orthodox probability theory, where the 'probs' are measures over sets, these relationships are trivial consequences of the properties of those sets. However the fact that an algebra of the plausibilities of propositions has to obey the same rules, or else be inconsistent, is not at all obvious and Cox's demonstration of this can be accounted "the most important advance in the conceptual formulation of probability theory since Laplace" (Jaynes, in preparation).

Cox's assumptions or axioms were very simple:

1. The plausibility of $\overline{\mathcal{A}}$ should be determinable from the plausibility of $\mathcal{A}$;
2. The plausibility of $\mathcal{A}\cdot\mathcal{B}$ should be determinable from two other plausibilities, one that of $\mathcal{A}$ and the other of $\mathcal{B}$ given the truth of $\mathcal{A}$.[4]

Since we want plausibilities to be rankable, it is natural to assign them to be real numbers; this was an implicit assumption of Cox. Remarkably, this is all that is required to construct the algebra. There is always the risk that there might be a lurking inconsistency in the axioms, but none has yet been discovered. If the axioms are rewritten in symbols, then

$$p(\overline{\mathcal{A}}|\mathcal{C}) = f\big(p(\mathcal{A}|\mathcal{C})\big) \quad \text{and} \tag{1.3}$$

$$p(\mathcal{A}\cdot\mathcal{B}|\mathcal{C}) = F\big(p(\mathcal{A}|\mathcal{C}), p(\mathcal{B}|\mathcal{A}\cdot\mathcal{C})\big), \tag{1.4}$$

where f and F are two functions to be determined. Fortunately, properties of the propositional logic allow restrictions to be placed on these functions. For example, the plausibility $p(\mathcal{A}\cdot\mathcal{B}\cdot\mathcal{C}|\mathcal{D})$ can be worked out in many different ways using (1.4), because the proposition $\mathcal{A}\cdot\mathcal{B}\cdot\mathcal{C}$ can be written equivalently as, for example, $\mathcal{A}\cdot(\mathcal{B}\cdot\mathcal{C})$ and $(\mathcal{A}\cdot\mathcal{B})\cdot\mathcal{C}$, by using the associative property of the conjunction operation. In order to be consistent, all these evaluations of the same plausibility must have the same value. This implies that F satisfies the functional equation

$$F\big(x, F(y,z)\big) = F\big(F(x,y), z\big). \tag{1.5}$$

The solution of this equation is the product rule (1.2). Similarly, since $\overline{\overline{\mathcal{A}}} = \mathcal{A}$, the function f is required to satisfy

$$f\big(f(x)\big) = x. \tag{1.6}$$

A somewhat more intricate argument shows that, in order to be consistent with the product rule, f must also satisfy

$$xf\left(\frac{f(y)}{x}\right) = yf\left(\frac{f(x)}{y}\right). \tag{1.7}$$

---

[4] Because $\mathcal{A}\cdot\mathcal{B} = \mathcal{B}\cdot\mathcal{A}$, this axiom can be rephrased in terms of the plausibilities of $\mathcal{B}$ and of $\mathcal{A}$ given $\mathcal{B}$. Tribus (1969) has shown that, if the plausibility of $\mathcal{A}\cdot\mathcal{B}$ were to depend on any other quantities, absurdities would result.

These two relations lead to the sum rule (1.1). In addition, if $\mathcal{A}$ is certainly true (false), given $\mathcal{B}$, it has plausibility one (zero). Now that plausibilities have been shown to obey the same rules as probabilities, we can safely replace $p(\mathcal{A}|\mathcal{B})$ by $\mathrm{prob}(\mathcal{A}|\mathcal{B})$. However, it should be understood that, in the Bayesian usage, a probability is not defined in terms of repeated sampling. Rather it is an objective measure of the likely truth of a proposition in the light of some evidence, and, as such, of much wider applicability.

1.2.2 Developing the algebra

From the product rule (1.2), we can trivially derive the most important rule of probability theory, namely, **Bayes' theorem**. For, by exploiting the commutativity of the conjunction operation, namely that $\mathrm{prob}(\mathcal{A}\cdot\mathcal{B}|\mathcal{C}) = \mathrm{prob}(\mathcal{B}\cdot\mathcal{A}|\mathcal{C})$, we find that

$$\mathrm{prob}(\mathcal{A}|\mathcal{B}\cdot\mathcal{C}) = \frac{\mathrm{prob}(\mathcal{A}|\mathcal{C})\,\mathrm{prob}(\mathcal{B}|\mathcal{A}\cdot\mathcal{C})}{\mathrm{prob}(\mathcal{B}|\mathcal{C})}. \tag{1.8}$$

This is the rule which tells us how to update probabilities in the light of new information. Let $\mathcal{A}$ be the proposition which we are seeking to ascertain the truth of, that is, the theory, and $\mathcal{C}$ any information we have that is relevant to $\mathcal{A}$, and which leads us to assigning the probability of the theory to be $\mathrm{prob}(\mathcal{A}|\mathcal{C})$. We obtain new data described by $\mathcal{B}$ and want to revise the probability of $\mathcal{A}$ in the light of it. Bayes' theorem tells us to combine three probabilities in order to accomplish this: the first is $\mathrm{prob}(\mathcal{A}|\mathcal{C})$, which is called the **prior** probability and encodes what we know about $\mathcal{A}$ before the data is taken into account; the second is the probability of obtaining the measured data $\mathcal{B}$, assuming that the theory (and all the background information $\mathcal{C}$) were true, which is called the **likelihood** of the theory;[5] and the third is the probability that we would have got the data whether or not the theory were true, the **evidence**. The revised probability of $\mathcal{A}$, $\mathrm{prob}(\mathcal{A}|\mathcal{B}\cdot\mathcal{C})$ is called the **posterior**[6] probability distribution. Schematically,

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}. \tag{1.9}$$

We have not yet written down an expression for the probability of the disjunction of two propositions. That gap is filled now:

$$\mathrm{prob}(\mathcal{A}+\mathcal{B}|\mathcal{C}) = \mathrm{prob}(\mathcal{A}|\mathcal{C}) + \mathrm{prob}(\mathcal{B}|\mathcal{C}) - \mathrm{prob}(\mathcal{A}\cdot\mathcal{B}|\mathcal{C}), \tag{1.10}$$

which can be shown using the sum and product rules and De Morgan's law $\mathcal{A}+\mathcal{B} = \overline{\overline{\mathcal{A}}\cdot\overline{\mathcal{B}}}$. If the propositions $\mathcal{A}$ and $\mathcal{B}$ are **exclusive** given $\mathcal{C}$, that

[5] This is the modern Bayesian definition of the likelihood. Occasionally, one comes across the alternative definition $\mathrm{prob}(\mathcal{B}|\mathcal{A}\cdot\mathcal{C})/\,\mathrm{prob}(\mathcal{B}|\mathcal{C})$.

[6] 'Prior' and 'posterior' are often used as nouns, with 'probability' implied.

is, if they cannot simultaneously be true, then this expression simplifies to $\mathrm{prob}(\mathcal{A}+\mathcal{B}|\mathcal{C}) = \mathrm{prob}(\mathcal{A}|\mathcal{C}) + \mathrm{prob}(\mathcal{B}|\mathcal{C})$. In a set of N exclusive propositions $\{\mathcal{A}_n\}$, only one proposition can be true. That is, $\mathrm{prob}(\mathcal{A}_n \cdot \mathcal{A}_m|\mathcal{B}) = 0$ if $m \neq n$. The probability of the disjunction of the elements of this set $\mathrm{prob}(\mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_N|\mathcal{B})$ equals $\sum_n \mathrm{prob}(\mathcal{A}_n|\mathcal{B})$. A set of propositions is **exhaustive** if at least one element of that set is true, that is, if $\mathrm{prob}(\mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_N|\mathcal{B}) = 1$. In this thesis, we shall be entirely concerned with sets of propositions which are both exclusive and exhaustive. In such sets, one and only one proposition is supposed true.

The term **hypothesis space** is one that will crop up later in this thesis. This is the set of all propositions, usually exclusive, and, by *fiat*, exhaustive, that we are considering during one particular probabilistic calculation. As we learn, this space of possibilities will change. It might enlarge, for example, if we choose to add new parameters to a model; on the other hand, it might shrink, if we manage to unify two theories by showing that their parameters are related. Alternatively, if we are on the wrong track, a completely different hypothesis space might be needed. The creative part of science is the concoction of new hypothesis spaces and there are very few general rules to help in the search.

As well as being able to update probabilities, it turns out to be very useful to have a technique for eliminating uninteresting propositions from them. Suppose we have constructed the set of probabilities $\mathrm{prob}(\mathcal{A} \cdot \mathcal{B}_n|\mathcal{C})$ for some $\mathcal{A}$ and an exclusive and exhaustive set $\{\mathcal{B}_n\}$, but that it is only $\mathcal{A}$'s probability we want to work out. The $\mathcal{B}$s might assert the value of a nuisance parameter, such as an unknown noise level. Now, by exhaustivity, the proposition $\mathcal{B}_1 + \cdots + \mathcal{B}_N$ is true and hence $\mathcal{A} = \mathcal{A} \cdot (\mathcal{B}_1 + \cdots + \mathcal{B}_N) = \mathcal{A} \cdot \mathcal{B}_1 + \cdots + \mathcal{A} \cdot \mathcal{B}_N$, and $\mathrm{prob}(\mathcal{A}|\mathcal{C}) = \mathrm{prob}(\mathcal{A} \cdot \mathcal{B}_1 + \cdots + \mathcal{A} \cdot \mathcal{B}_N|\mathcal{C})$. But, since the $\mathcal{B}$s are exclusive, so are the $\mathcal{A} \cdot \mathcal{B}_n$ and we obtain the result

$$\mathrm{prob}(\mathcal{A}|\mathcal{C}) = \sum_n \mathrm{prob}(\mathcal{A} \cdot \mathcal{B}_n|\mathcal{C}). \tag{1.11}$$

This is the **marginalization** rule: we speak of marginalizing away the $\mathcal{B}_n$. As a bonus, it allows us to evaluate the term in Bayes' theorem which is called the evidence. For $\mathcal{A}$ and $\overline{\mathcal{A}}$ are certainly exhaustive and exclusive propositions. Hence in (1.8), $\mathrm{prob}(\mathcal{B}|\mathcal{C}) = \mathrm{prob}(\mathcal{B} \cdot (\mathcal{A} + \overline{\mathcal{A}})|\mathcal{C}) = \mathrm{prob}(\mathcal{B} \cdot \mathcal{A}|\mathcal{C}) + \mathrm{prob}(\mathcal{B} \cdot \overline{\mathcal{A}}|\mathcal{C})$. But the first term is just the numerator on the RHS of (1.8) and the second term is the numerator in an application of Bayes' theorem to update the negated proposition $\overline{\mathcal{A}}$. In order to be able to evaluate the terms in Bayes' theorem applied to $\overline{\mathcal{A}}$, though, it is necessary to have a completely specified hypothesis space. If that space consists of the exclusive and exhaustive set $\{\mathcal{A}_n\}$, then to assert that $\mathcal{A}_m$ is false is to say that one of the other propositions is true. We cannot apply Bayes' theorem if the other $\mathcal{A}_n$ are not specified because $\mathrm{prob}(\mathcal{B}|\mathcal{A}_n \cdot \mathcal{C})$ cannot be determined (Bretthorst 1988).

The results just discussed hold for an arbitrary number of discrete propositions. Often however we want to reason about a quantity which can take on any value in a continuum; the propositions then have a continuous label and it is most useful to introduce a **probability density function**. Let $\mathcal{A}(x)$ assert that some quantity has a value x. Then the probability that the quantity takes on a value between x and $x+\mathrm{d}x$, $\mathrm{prob}([\mathcal{A}(x), \mathcal{A}(x+\mathrm{d}x)]|\mathcal{B})$, is written as $\mathrm{Prob}(\mathcal{A}(x)|\mathcal{B})\,\mathrm{d}x$, where Prob is the density function. Fortunately, Bayes' theorem is still true for Prob functions. If the propositions $A(x)$ are exhaustive and exclusive, then $\mathrm{Prob}(\mathcal{A}(x)|\mathcal{B})$ must satisfy the normalization condition $\int \mathrm{Prob}(\mathcal{A}(x)|\mathcal{B})\,\mathrm{d}x = 1$. In addition, sums over discrete probabilities, such as appear in the marginalization rule (1.11), are replaced by integrals over probability densities.

1.2.3 Jaynes' consistency principle

So far we have not talked at all about assigning probabilities. Rather it has been shown how they can be manipulated once they have been obtained. In order to keep the theory objective, it is necessary that people with the same information about a problem reason in the same way. This is the content of Jaynes' consistency principle (Jaynes 1968), which states that equivalent states of knowledge must give rise to equivalent probability assignments. This applies to all probabilities, both priors and likelihoods. From this principle at least two important techniques for assigning probabilities spring. One is the principle of maximum entropy, which is discussed in the next section. Secondly, invariance arguments can be used. The idea here is that the hypothesis space can be reparametrized in some way that leaves our state of knowledge about the parameters unchanged. The consistency principle means that the same probabilities must be assigned to the new propositions as to the old ones. But the relation between the old and the new propositions is known and this restricts the possible probability assignments. It is the invariance of our information under the reparametrization of the system that is the key to this approach. Usually the only cases where such an invariance is identifiable is when we have a very little general knowledge about the propositions and we want to assign **uninformative** priors. For example, the flat prior distribution for a locational parameter and the Jeffreys prior for a scale parameter can be derived in this way (Jaynes 1968). I shall use this technique in Chapter 4, when a prior for a spectral density function will be sought.

1.3 Entropy and probability assignment

1.3.1 The principle of maximum entropy

That we have to assign probabilities to propositions rather than reason deductively is a sign that we do not have complete knowledge about a situation; uncertainty is present. It is possible uniquely to construct a numerical measure of the uncertainty that is implicit in a probability assignment to a set of propositions. It takes a very simple form when those propositions are exhaustive and exclusive:

$$S = -\sum_n p_n \ln p_n, \tag{1.12}$$

where $p_n = \text{prob}(\mathcal{A}_n|\mathcal{B})$. This quantity is called the **entropy** of a probability distribution. It is maximized when all the probabilities are equal and it then has the value $\ln N$. On the other hand, it has its minimum value of zero when one proposition is certainly true. This corresponds nicely with the intuitive idea of uncertainty. It was Shannon (1948) who discovered that the entropy formula was a (negative) measure of information—in modern Bayesian terms, a measure of uncertainty.

The entropy can be shown to be the only possible measure consistent with a set of plausible axioms:

1. The measure should be a continuous function of the probabilities;
2. If the propositions are equiprobable, the uncertainty should increase as a monotonic function of their number;
3. Different ways of determining the uncertainty inherent in the probabilities of a given set of propositions should give the same result.

It was proposed by Jaynes (1957) that this function had an important part to play in the assignment of probabilities. Given a certain kind of information, called **testable**, which is interpretable directly in terms of the probabilities of a set of propositions and restricts their possible values, although not enough to isolate one single probability assignment, does it not correspond nicely with intuition that one should choose that distribution which, while satisfying the constraints, is as non-committal as possible, that is, has the greatest uncertainty? This method has come to be called the **principle of maximum entropy**: it is not to be confused with the 'Maximum Entropy Method' or MaxEnt. Other axiomatic approaches have confirmed this to be a valid way to assign probabilities (e.g., Shore and Johnson 1980). The only slight difficulty in its use is to identify what sort of physical measurements can be considered as testable information. In this regard, the theory is still incomplete, but a crucial component is that the measurement, although it may vary from one experiment to the next, should have some reproducible features. For example, a body in equilibrium and in contact with a large heat bath has a reproducible average energy, which can be used as testable information to assign the probabilities of the occurrence of the energy eigenstates of the body.[7]

[7] It was Jaynes' recognition (1957) that Gibbsian statistical mechanics was really a theory of the assignment of probabilities by maximizing entropy which set the modern Bayesian band-wagon in motion.

When the propositions form a continuum, it is necessary to think again about the form of the entropy. The obvious generalization of (1.12)—the replacement of the sum by an integral—is flawed since the resulting entropy is not invariant under a change of the variable labelling the propositions. In fact, it is necessary to introduce a function $m(x)$ into the expression, which is the probability assignment of maximal entropy in the absence of any data:

$$S = -\int p(x) \ln\left(\frac{p(x)}{m(x)}\right) \mathrm{d}x, \tag{1.13}$$

where $p(x) = \mathrm{Prob}(\mathcal{A}(x)|\mathcal{B})$ (Jaynes 1963).

1.3.2 Gaussian noise

One of the simplest applications of the principle of maximum entropy is to assigning a probability density distribution to the noise component of a measurement of some quantity. Here it is supposed that the noise has a reproducible average value of zero and some reproducible variance σ^2. These are pieces of testable information relevant to $p(x)$, the probability that the noise has the value x:

$$0 = \int_{-\infty}^{\infty} x p(x)\, \mathrm{d}x \qquad \text{and} \tag{1.14}$$

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p(x)\, \mathrm{d}x. \tag{1.15}$$

The only assignment of $p(x)$ which does not implicitly assume something else about the noise is the one which maximizes the entropy (1.13), subject to the two testable constraints and normalization. If we assign a flat default function, we obtain for the probability distribution of the noise

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right), \tag{1.16}$$

which is a Gaussian distribution. If similar testable information is available for the noise components of N data, another entropic argument shows that the set of noise components $x_1, \ldots, x_N$ should be assigned a multivariate Gaussian distribution independent in the x_n:

$$p(x_1, \ldots, x_N) = \prod_n \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{x_n^2}{2\sigma_n^2}\right), \tag{1.17}$$

where the nth noise variance is σ_n^2. This noise distribution will be used throughout this thesis.

1.3.3 Entropy vs configurational entropy

We shall see in Chapter 2 that the entropy expression plays another important rôle in Bayesian calculations. It is shown there that a very useful prior for reconstructing positive images is proportional to the exponential of a constant multiplied by the *configurational* entropy of the image, that is, the entropy function applied to the image strengths $\{f_n\}$ rather than to a probability distribution $\{p_n\}$.[8] It is this prior which leads to the 'Maximum Entropy Method' of image reconstruction, now more often called, simply, MaxEnt. It is tempting to seek a unifying principle that connects the entropy of a probability distribution, which is a measure of uncertainty, to the entropy of an image, which is a measure of prior probability. However the connection remains obscure. Until a link is found, it is the author's opinion that confusion is best avoided if the ideas are kept firmly distinct, and remarks to the effect that probabilities are like images or *vice versa* are taken with a pinch of salt.

1.4 The Bayesian revolution

In the last fifteen years, the number of successful applications of Bayesian methods has increased enormously, as people have begun to realize that this rather abstract theory can produce orders of magnitude improvements in problems of inference. Often the computational aspects are extremely challenging, however. For hypothesis spaces with a large number of parameters, it is necessary to tame difficult searches and integrations over probability densities of an enormous number of variables. It is no doubt partly for this reason that the Bayesian approach has been in eclipse for most of this century. The modern flowering would not have been possible without the digital computer. Had Jeffreys (1939), one of very few Bayesians writing at that date, been able to illustrate his remarkable book with some of today's examples of image reconstruction, for example, his lone voice might have been listened to more sympathetically.

Two highlights, where the difficulties have been largely overcome, have been the development, in Cambridge, England, of the theory and the algorithms to reconstruct large images (e.g., Skilling 1989; Gull 1989)—this thesis is concerned with developments and applications of that theory—and work, at Washington University, St. Louis, which has demonstrated the principles and power of Bayesian parameter estimation and model comparison (e.g., Bretthorst 1988).

---

[8] It is related to the number of ways (configurations) in which the image can be built up: see the 'monkey argument' of Chapter 2.

1.5 Conclusion

This chapter has introduced the Bayesian probability theory. It has been necessarily brief, but more detailed reviews are available elsewhere. Particularly to be recommended are the book of Jaynes, which is still incomplete, although sections are available from the author (Jaynes, in preparation), and a recent review by Loredo (1990). In the chapters to follow, I shall put this theory to work on solving inverse problems and estimating spectral density functions.

References

Bretthorst, G. L. (1988). *Bayesian spectral analysis and parameter estimation* (Lecture notes in statistics, Vol. 48). Springer-Verlag, New York.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, **14**, 1–13. Later expanded into Cox (1961).

Cox, R. T. (1961). *The algebra of probable inference.* Johns Hopkins Press, Baltimore.

Gull, S. J. (1989). Developments in maximum entropy data analysis. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 53–71. Kluwer, Dordrecht.

Jaynes, E. T. (1957). Information and statistical mechanics. *Physical Review*, **106**, 620–30.

Jaynes, E. T. (1963). Information theory and statistical mechanics. In *Statistical physics,* 1962 Brandeis Lectures (ed. K. W. Ford), pp. 181–218. Benjamin, New York.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, **SSC-4**, Sept., 227–41.

Jaynes, E. T. (in preparation). *Probability theory: the logic of science.* Fragments available from the author: Wayman Crow Professor of Physics, Washington University, St. Louis, MO 63130, USA.

Jeffreys, H. (1939). *Theory of probability.* Clarendon Press, Oxford. Later editions: 1948 and 1961.

Loredo, T. J. (1990). From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In *Maximum entropy and Bayesian methods, Dartmouth, U.S.A., 1989* (ed. P. F. Fougère), pp. 81–142. Kluwer, Dordrecht.

Polya, G. (1954). *Mathematics and plausible reasoning. Induction and analogy in mathematics* (Vol. 1); *Patterns of plausible inference* (Vol. 2). Princeton University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423 and 623–59. Reprinted in C. E. Shannon and W. Weaver (1949). *The mathematical theory of communication.* University of Illinois Press, Urbana.

Shore, J. E. and Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, **IT-26**, 26–37.

Skilling, J. (1989). Classic maximum entropy. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 45–52. Kluwer, Dordrecht.

Tribus, M. (1969). *Rational descriptions, decisions and designs.* Pergamon, New York.

2

Inverse problems and the search for priors

The game was Rouge et Noir. I had played at it in every city in Europe, without, however, the care or the wish to study the Theory of Chances— that philosopher's stone of all gamblers.
Wilkie Collins: A Terribly Strange Bed

In this chapter, I shall apply the very general algebra of probable inference which was described in Chapter 1 to the problem of inferring density functions from incomplete and noisy measurements of some integral transform linear in the density. I shall discuss the rôle of the prior in such problems and, in particular, describe the Gaussian and entropic forms.

2.1 What is an inverse problem?

It is often the case that a physically interesting quantity is not accessible from direct measurement. Experimental data are most easily obtainable on some transform of this quantity, a quantity which we shall call, no matter what its precise physical nature, the **image**. It might be the power spectrum of a voiced vowel or the charge density of a ^{12}C nucleus in its ground state, for example. The action of the apparatus is modelled as an operator acting on the image, which it transforms in some way, before spitting out a datum. Then the **inverse problem** is that of inferring the image from the data. There are two broad aspects to this. The first we might call *deductionist*. It consists of the mathematical theory of transform inversion, showing under what conditions and exactly how transforms can be inverted. For instance, the Bromwich integral solves the inverse problem arising from the Laplace transform. Secondly, and much more central to the interpretation of experiments, there is *probabilistic* inversion. This theory seeks to show how to infer plausible images from necessarily imperfect and incomplete experimental data. It turns out that these two questions require very different methodologies and we shall not be considering the first at all.[1]

[1] This contrasts with many orthodox statistical approaches to the second problem which start from solutions to the first and proceed to modify them in a plausible but

The qualities of imperfection and incompleteness of data merit some further consideration. The first refers to the fact that there is nearly always a significant extra contribution to the measured data which is not completely accounted for by the operator modelling the apparatus and which it is not possible or practical to describe precisely: this is experimental noise. In principle, it could be described fully and incorporated into the operator. That is, there is nothing intrinsically nondeterministic about it. But the extra information that would be necessary to be able to do this is not available and we are satisfied to describe what we do know about the noise by means of a probability distribution. An analogy with heating and performing work in thermodynamics is useful. Working represents the perturbation of a system by a *known* change in the Hamiltonian and parallels the assumed known image-to-data transform. Heating, although still a change in the Hamiltonian, represents a temporary *unknown* change and is described probabilistically: this is the analogue of noise. It is clear that the borderline between noise and signal is rather blurred and indeed one man's noise can be another man's signal.

Even with no noise present, however, the inverse problem would, in general, still not be soluble using deductionist approaches. For incompleteness rears its ugly head. This means that there is not enough information in the data set to allow for the use of an inverse transform. For instance, in a scattering problem, the apparatus might not allow us to measure a cross-section at a high enough momentum-transfer to be able to deduce the high frequency components of a charge distribution (see Chapter 3). Again it is necessary to describe in a probability distribution what we know, in this case about the unmeasured degrees of freedom.

As we saw in Chapter 1, Bayes' theorem is the tool which we use to learn from new evidence, and in the following sections we shall consider how to assign the probabilities that appear in that theorem. First however we narrow the field even more.

An important class of inverse problems is the one with integral operators acting linearly on the image, plus additive noise. Let us establish a notation. The image we shall denote by $f(x)$ with $x \in \mathbb{R}^\nu$. The pth of P data is D_p and the error on that datum is n_p. Then

$$D_p = \int \hat{O}(p,x) f(x) \, \mathrm{d}^\nu x + n_p. \tag{2.1}$$

The function $\hat{O}$ is the response function which describes how much of the image at each x is carried into the pth datum. In practice, it usually becomes necessary to 'bin' the image, that is, approximate the integral

---

ad hoc manner to obtain 'solutions' to the second. The usual discussion of the discrete Fourier transform as a spectral estimator falls into this category.

as a sum, especially at the computational stage. The image $f(x)$ then becomes an N-element column matrix $\boldsymbol{f}$, say, and $\hat{O}$ a $P \times N$ matrix $\mathbf{O}$. In the sequel, I shall slip between this integral/sum notation and one using kets which more clearly displays whether operators are acting on images or data vectors. The image is then $\mid f\rangle$ and the response for the pth datum $\mid p\rangle$. If $\mid n\rangle$ is the ket corresponding to the image being entirely in the nth bin of the discretized image, then f_n is equivalent to $\langle n \mid f\rangle$, $\langle n \mid p\rangle = O(p, x_n)\Delta x$, where Δx is the bin width, and in this discrete approximation, the identity operator is $\sum_n \mid n\rangle\langle n \mid$. We can write (2.1) more compactly as $D_p = \langle p \mid f\rangle + n_p$. The real vector space in which images live, $\mathcal{I}$ say, is called **image space** while data vectors lie, not surprisingly, in **data space**.

2.2 The assignment of terms in Bayes' theorem

2.2.1 The Gaussian likelihood

In Chapter 1, we saw how to motivate assigning uncorrelated Gaussian distributions to the noise terms $\{n_p\}$. Let us take this as given and see what emerges. In that case, the assignment of the likelihood is straightforward. For, given the image f and the relation between that image and data (2.1), the probability of the data given the image is just the probability that the noise makes up the difference $D_p - \langle p \mid f\rangle$ for each p. That is

$$\mathrm{prob}(\boldsymbol{D}|\boldsymbol{f} \text{ and } \mathcal{K}) = \frac{1}{\prod_p \sqrt{2\pi}\sigma_p} \exp\left(-\sum_{p=1}^{P} \frac{\left(D_p - \langle p \mid f\rangle\right)^2}{2\sigma_p^2}\right). \tag{2.2}$$

Here σ_p is the width of the Gaussian assigned to the error probability of datum p. The proposition $\mathcal{K}$ asserts all the background information that goes into the model for the signal and the noise, of which only a small part is relevant to this probability.

2.2.2 Modelling the image

At this point it is important to consider just what mathematical description for the image we choose to use. This is the problem of choosing the hypothesis space. We consider two broad categories (cf. Daniell 1991). The first supposes that we can construct a model with a small number of parameters to describe the image. By small we mean small relative to the amount of data at hand. The model might have arisen out of some theory of the image, a theory which leaves parameters undetermined. Then the inverse problem is to infer these parameters. On the other hand, we might have no obvious model and just want to do the best we can at estimating the whole of $f(x)$, the image. This is a much more demanding problem because the number of variables describing the image is much larger in general than the number of data.[2] We cannot expect the data to determine the image sharply by itself: the rôle of the prior becomes crucial.

These broad classes lie on a spectrum of possibilities measured by the number of parameters to be estimated relative to the amount of data. Although it is not easy to separate them unambiguously and indeed they shade off into each other as the number of variables approaches the number of independent data points, the classification is still useful, for each category leads to somewhat different, and complementary, approaches to inverse problems. The first, which we shall refer to as **Bayesian parameter estimation**, has been pioneered by Jaynes (1987) and Bretthorst (1988), following on from work by Jeffreys (1939), the second, **free-form modelling**, by Daniell, Gull and Skilling (Gull and Daniell 1978; Skilling 1989; Gull 1989). Power in solving inverse problems comes from the ability to exploit both. The free-form method, unhindered by a specific model of the image, does not restrict the range of possible reconstructions. Thus it is good for the exploratory phase of an analysis when more specialized models are not available. The resulting free-form images can be used as very effective guides to such models. Examples of this style of approach appear, for example, in the analysis of sunspots in Chapter 5, and in the work of Sibisi (1990).

2.2.3 The structure of the likelihood

Now let us look at the likelihood in the case of an $N > P$ problem. The exponent in the likelihood (2.2) multiplies out to

$$\text{constant} - \tfrac{1}{2}\left(\langle f \mid \hat{K} \mid f \rangle - 2 \sum_p \frac{D_p}{\sigma_p^2} \langle p \mid f \rangle \right), \tag{2.3}$$

where the operator $\hat{K}$, which we call the **kernel**, satisfies

$$\hat{K} = \sum_p \frac{\mid p \rangle \langle p \mid}{\sigma_p^2}. \tag{2.4}$$

Its eigenstructure fixes the directions in image space which are determined by the data alone, as follows. Being a symmetric operator it can be expressed in the diagonal form

$$\hat{K} = \sum_q \mid g_q \rangle \Lambda_q^2 \langle g_q \mid, \tag{2.5}$$

[2] That $f(x)$ is a *function*, and hence has an infinite number of 'parameters', is not a problem in real applications, for we can adequately represent it by its values at selected x_i. There will often be a very large number of these, however, and this is where the difficulty arises.

where $| g_q\rangle$ is an eigenket of $\hat{K}$ with eigenvalue Λ_q^2, which is necessarily non-negative because of the special, projective, form of $\hat{K}$. These eigenkets are images which are inferentially independent in the light of the data. It is easy to show that at most P of the N Λ_q^2 are non-zero. Image space is severed into two distinct subspaces ($\mathcal{I} = \mathcal{N} \oplus \mathcal{M}$), one of zero eigenvalues (the **null space** $\mathcal{N}$) which the data can do nothing to help us chart. The amounts of the image made from eigenvectors in the other, **measured**, subspace ($\mathcal{M}$) are determined with varying uncertainties, large eigenvalues corresponding to well-measured directions, smaller eigenvalues to less well-measured ones.[3] It is rare that all P directions in $\mathcal{M}$ are well-determined: some data are badly measured, while some are redundant, which is to say, the same kind of information about the image is present in more than one data point. As a result there are less than P good independent measurements. The kernel projects away any part of an image in $\mathcal{N}$ and weights, via σ_p^{-2}, badly-measured directions in the rest of $\mathcal{I}$ more weakly than well-measured ones. Here, at least, one vital rôle of the prior is clear: it will have to extrapolate into $\mathcal{N}$. The theoretical discussion in Chapter 1 showed why a prior appears in Bayes' theorem; here we see a pragmatic reason for its presence.

The linear algebra of the kernel is especially important for future discussions so I shall summarize it here. The eigenkets can be, and are assumed to have been, orthonormalized: $\langle g_{q'} \mid g_q\rangle = \delta_{q'q}$. The identity operator in $\mathcal{I}$ is then represented as $1_{\mathcal{I}} = \sum_{q\in\mathcal{I}} | g_q\rangle\langle g_q |$,[4] while in $\mathcal{M}$, the identity is made just by summing over the qs labelling vectors with non-zero eigenvalue. We can express the response function in the measured space in terms of the $| g_q\rangle$s—this is the so-called singular value decomposition (see, for example, Press *et al.* 1986):

$$\frac{| p\rangle}{\sigma_p} = \sum_{q\in\mathcal{M}} | g_q\rangle \frac{\langle g_q \mid p\rangle}{\sigma_p} = \sum_{q\in\mathcal{M}} | g_q\rangle \Lambda_q V_{pq}. \tag{2.6}$$

The last term V_{pq} in the final equality equals $\langle g_q \mid p\rangle/(\sigma_p\Lambda_q)$ and has a special significance. It is not hard to show that it is the pth component of the qth normalized eigenvector of a $P \times P$ matrix with elements $\kappa_{pp'} = \langle p \mid p'\rangle/(\sigma_p\sigma_{p'})$, which we call *dual* to the kernel. The eigenvalue is Λ_q^2 and so the whole vector is a data space counterpart to $| g_q\rangle$. Each Λ_q is the positive square-root of Λ_q^2. Since this dual matrix is symmetric, the matrix $\mathbf{V}$ is orthogonal, that is, $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_P = \mathbf{V}\mathbf{V}^{\mathrm{T}}$.

[3] Of course, it is not the directions or eigenvectors which are well or badly measured, but the coefficients of those vectors in an expansion of the image in terms of the eigenvectors. This is what I shall mean by well- or badly-measured directions.

[4] The notation $q \in \mathcal{I}$ is somewhat metaphorical. It means not that it is q that is in $\mathcal{I}$ but rather the eigenket labelled by q.

Because of the independence property noted above, it is convenient to describe the image with the eigenkets of the kernel.[5] It is then straightforward to show that the image which maximizes the likelihood and has no null component is

$$| f_{\text{datagram}}\rangle = \sum_{q\in\mathcal{M}} D'_q | g_q\rangle, \tag{2.7}$$

where

$$D'_q = \sum_p \frac{D_p}{\sigma_p}\frac{V_{pq}}{\Lambda_q}. \tag{2.8}$$

In the case where the transform kets are orthonormal to start with,[6] this image reduces to $\sum_p D_p | p\rangle$. This ket is called the **datagram** because it is an image which is entirely equivalent to the data. In the spectral problem of Chapter 4, it is closely related to the Schuster periodogram.

However as a solution to the inverse problem, it is lamentable. Assigning the null component of the image to zero is usually a very bad extrapolation and leads to spurious truncation features in the reconstruction. Similarly, fitting the data exactly, as occurs with a datagram in an $N > P$ problem, means that there is no suppression of structure arising from noise. But most damning of all, it is impossible to quantify how reliable is this one image, for we have not yet got to the stage of constructing a probability distribution for the image given the data. We have been looking just at a distribution which describes how uncertain we are of the data given the image. We should have been suspicious when we saw (2.8), for the appearance of Λ_q in the denominator forebodes indeterminacy if a direction were to become null, that is, if its $\Lambda_q \to 0$.

2.3 What to use for a free-form prior?

The prior distribution is meant to encode all the information that we have about the image before the data is taken into account. Usually this information is very complex and differs widely from one problem to the next. In addition, it is often of such a qualitative nature that it is quite difficult to formulate in a way precise enough to describe it in a prior. Indeed this is an area where much more research needs to be done. However if the data is good enough, it is often satisfactory to neglect any pre-data knowledge specific to a particular problem. In that case, we want to be able to take a general purpose prior out of our tool-bag and start chipping at the data with that. This is a safe way to start: if the answer confounds expectations, it means one had prior knowledge which really *was* essential. At that point, we would have to think about incorporating it.

[5] Or its close relation. When a prior is included it will, in general, modify the measured space.

[6] We shall see such a case when we tackle spectral estimation in Chapter 4.

So what constitutes a general purpose prior? We look at two such functions in the next section and investigate their properties.

2.3.1 The Gaussian prior: a soluble case

Occasionally we might have testable information on the image, which it is easy to incorporate using Jaynes' principle. One such piece of information might arise from a reproducible macrovariable, quadratic in the image, an image 'power'. This was the kind of information we assumed about the noise when we derived for it a Gaussian distribution in Chapter 1. For example, we might have reason to suppose that a particular time-series is stationary with regard to its power over different intervals and at different times—stationary, that is, in the sense that the variable concerned is quite reliably constant. This can be applied as testable information on the spectrum of the series. The constraint might be $\sum_n f_n^2 = X$ and the usual formalism (maximizing the entropy of the probability distribution of the f_ns subject to this constraint) leads to a prior which is a Gaussian function of the image:

$$\mathrm{prob}(\{f_n\}) \propto \exp\Big(-\tfrac{1}{2}\alpha \sum_n f_n^2\Big), \tag{2.9}$$

where $\alpha > 0$ is a constant, determined in principle from X. When this constant is large, the zero image is strongly favoured *a priori*. There is a quadratic penalty in the logarithm of the prior (the log-prior) for an image not being zero. This prior will also tend to smooth the image. Suppose we had a two cell image (f_1, f_2) on which there was data that was good enough only to resolve the total intensity in the image $f_1 + f_2$, but not the individual fs.[7] This would correspond to the likelihood being very sharply peaked in the $f_1 + f_2$ direction and very flat in the $f_1 - f_2$ direction, say. Noise could easily move the image about in this 'slack' direction, so that predictions made by maximizing the likelihood of f_1 and f_2 would be very unreliable; although the data is blind to $f_1 - f_2$, it might well be inferred to be some non-zero value. However, when the above prior is used, things change. Let us idealize and suppose the data fixes $f_1 + f_2$ to equal f, say. Then the log-prior penalty is proportional to $f_1^2 + f_2^2$, which can be rewritten as $f_1^2 + (f - f_1)^2$. This is minimized when $f_1 = f_2 = \frac{1}{2}f$, which is the uniform image. We say that the prior has **regularized** the badly-behaved likelihood. The penalty function or log-prior is a **regularizer**.

Using the identity $f_n = \langle n \mid f\rangle$ and the expression for unity in discrete image space, $1 = \sum_n \mid n\rangle\langle n \mid$, we can rewrite $\sum_n f_n^2$ as $\langle f \mid f\rangle$. The posterior probability of the image is obtained using Bayes' theorem:

[7] This is the kind of situation that arises in Fourier-like inversions (see Chapters 3 and 5) where the data only measures structure coarser than a certain scale in the image.

$$\begin{aligned} \text{prob(image}|\text{data)} &\propto \text{prior} \times \text{likelihood} \\ &\propto \exp\Big(-\tfrac{1}{2}\big(\alpha\langle f \mid f\rangle + \sum_p \left[D_p - \langle p \mid f\rangle\right]^2/\sigma_p^2\big)\Big). \end{aligned} \tag{2.10}$$

Because both the prior and the likelihood were Gaussian, so is the posterior. The argument of the exponential is

$$\text{constant} - \tfrac{1}{2}\Big(\langle f \mid (\alpha 1_{\mathcal{I}} + \hat{K}) \mid f\rangle - 2\sum_p \frac{D_p}{\sigma_p^2}\langle p \mid f\rangle\Big) \tag{2.11}$$

(cf. (2.3)). Luckily, the eigenstructure of the unregularized likelihood is still appropriate for this new problem. For the ket $\mid g_q\rangle$ is an eigenket of $\alpha 1_{\mathcal{I}} + \hat{K}$ of eigenvalue $\alpha + \Lambda_q^2$; the directions in image space about which we can make independent inferences are unchanged. Since $\Lambda_q \geq 0$ and $\alpha > 0$, there is no danger of $\alpha + \Lambda_q^2$ equaling zero: for this prior, an unmeasured direction has a probability with a finite width determined by α.

Using the eigenket expansion $\mid f\rangle = \sum_{q\in\mathcal{I}} c_q \mid g_q\rangle$, we find the most probable (and, in this case, also the mean) image to be

$$\mid f_0\rangle = \sum_q \overline{c_q} \mid g_q\rangle, \tag{2.12}$$

where

$$\overline{c_q} = \sum_p \frac{D_p}{\sigma_p}\frac{V_{pq}\Lambda_q}{\alpha + \Lambda_q^2}. \tag{2.13}$$

Note that, as one might have expected, these two expressions goes over into (2.7) and (2.8) as the regularizer becomes flat, that is, as $\alpha \to 0$. The difficulty, noted below (2.8), when a Λ_q goes to zero, is removed, because the corresponding coefficient goes gracefully to zero, also. However straightaway, a particular failing of this Gaussian prior is highlighted. The most probable coefficient of any unmeasured direction is zero. This means that there is no extrapolation outside of the measured space and into the null space. For problems where it is not reasonable to expect that the transform of the image outside the region where one has measured it is likely to be zero, this observation renders the present prior inappropriate. It is incapable of suppressing truncation artifacts, such as the sinc-wiggles that can be troublesome in spectral estimation problems.

Another problem is that the regularization is rather rudimentary. Consider a problem in which the response functions are orthonormal, that is, $\langle p_r \mid p_s\rangle = \delta_{rs}$ for all pairs of data points, labelled by r and s.[8] Suppose

[8] In Chapter 4, the response functions $\text{cas}(2\pi\nu r)$, with r an integer, will be used. They have this property when the scalar product is an integration over ν between $-\frac{1}{2}$ and $\frac{1}{2}$.

also that all the data points have the same error: $\sigma_r = \sigma\ \forall r$. In this case, each eigenket $|\, g_q\rangle$ reduces to $|\, p_q\rangle$ itself, the eigenvalue Λ_q is equal to σ^{-1} and **V** is simply the identity matrix. As a result, the most probable image is radically simplified, and becomes

$$|\, f_0\rangle = \frac{1}{\alpha\sigma^2 + 1} \sum_{p=1}^{P} D_p |\, p\rangle. \tag{2.14}$$

But this is just the datagram scaled by a factor $(\alpha\sigma^2+1)^{-1}$. Since $(\alpha\sigma^2 + 1)^{-1}$ is always less than or equal to one, there is no way to increase the intensity in any part of the image. If either the noise variance is very large or the regularizing strength α has been turned up strongly, so that $\alpha\sigma^2 \gg 1$, the tendency of the prior to pull the reconstruction to zero is felt keenly everywhere in the image.

The difficulty with this prior is that it penalizes the production of large intensity in the image too strongly. A regularizer with a weaker dependence on image strength is required. In the next section, one such function is discussed.

2.3.2 The entropic prior

For the inference of images that are known to be positive, a general purpose prior based on the configurational entropy of the image in a discretized 'bin' representation has been in use for a number of years (Gull and Daniell 1978). For images where the contents of a bin can be identified as the amount of something which it is meaningful to add up over bins, so-called **additive** images, and for which the only measurable quantities are sums linear in the amount of stuff in each bin, a demonstration that the entropic prior has a privileged status has been given by Skilling (1991). Other 'proofs' rely on this prior being the only one to satisfy certain plausible desiderata of consistent reasoning such as the monkey and kangaroo conditions (Shore and Johnson 1980; Gull and Skilling 1984; Jaynes 1986).

It is useful to rehearse a derivation of the modern form of this entropic prior, given first by Skilling (1989), for I shall use a similar idea to derive a new prior in Chapter 3. The idea is to invent a simple mechanism by which the image *could* have been generated, on the principle that a general purpose prior should apply in particular cases. We imagine a model in which a troupe of monkeys distributes an inexhaustible supply of small image elements of size ϵ into the discretized image of N bins. The image is specified by indicating how many elements end up in each of the bins, that is, the set of non-negative integers $\{n_i : i = 1 \to N\}$. Suppose that this is done in such a way that our state of knowledge of the outcome is described by a product of Poisson distributions, independent in all the bins.[9]

[9] This is a variation on the original monkey argument of Gull and Daniell (1978). In

The probability of a particular configuration is

$$\operatorname{prob}(\{n_i\}) = \prod_{i=1}^{N} \frac{\mathrm{e}^{-\nu_i}\nu_i^{n_i}}{n_i!} = \exp\Big(\sum_i \Big(-\nu_i + n_i \ln \nu_i - \ln n_i!\Big)\Big). \quad (2.15)$$

We suppose that in every bin there are sufficiently many elements that each $\ln n_i!$ can be approximated using Stirling's approximation, which can be written as $\ln n! \sim \frac{1}{2}\ln 2\pi + (n+\frac{1}{2})\ln n - n$, for n large. Inserting this into the above expression, gives

$$\operatorname{prob}(\{n_i\}) \simeq \frac{1}{\prod_i \sqrt{2\pi n_i}} \exp\Big(\sum_i \Big(n_i - \nu_i - n_i \ln \frac{n_i}{\nu_i}\Big)\Big). \quad (2.16)$$

The real image is presumably not quantized like the model, so we would like to construct from (2.16) a probability density for a continuous image. The configuration $(n_1, \ldots, n_N)$ corresponds to the image $(\epsilon n_1, \ldots, \epsilon n_N)$ which is the one image in this discrete model in a hypercube of image space with a volume ϵ^N centred on this point. We equate the probability density of the image multiplied by the volume of the hypercube to the probability of that particular configuration:

$$\epsilon^N \operatorname{prob}(\boldsymbol{f}) \simeq \frac{1}{\prod_i \sqrt{2\pi(f_i/\epsilon)}} \exp\Big(\frac{1}{\epsilon}\sum_i \Big(f_i - \nu_i\epsilon - f_i \ln \frac{f_i}{\epsilon\nu_i}\Big)\Big). \quad (2.17)$$

Simplifying this expression gives

$$\operatorname{prob}(\boldsymbol{f}) \simeq \frac{1}{\mathcal{Z}} \exp\big(\alpha S(\boldsymbol{f}, \boldsymbol{m})\big)\mu(\boldsymbol{f}), \quad (2.18)$$

where

$$S = \sum_n \left(f_n - m_n - f_n \ln \frac{f_n}{m_n}\right), \quad (2.19)$$

$$\mu(\boldsymbol{f}) = \prod_{n=1}^{N} f_n^{-1/2} \qquad \text{and} \quad (2.20)$$

$$\mathcal{Z} = \int_{\mathcal{I}} \exp\big(\alpha S(\boldsymbol{f}, \boldsymbol{m})\big)\mu(\boldsymbol{f})\, \mathrm{d}^N \boldsymbol{f}. \quad (2.21)$$

Each of the quantities m_n equals $\epsilon\nu_n$ and $\alpha = \epsilon^{-1}$. The function $S(\boldsymbol{f}, \boldsymbol{m})$ is called the **configurational entropy** of $\boldsymbol{f}$ with respect to the model $\boldsymbol{m}$.

---

that scenario, the monkeys had a fixed number M of image elements, the knowledge of the resulting configuration of which was described with the multinomial distribution $\operatorname{prob}(\{n_i\}) = N^{-M} M!/\prod_i n_i!$. This leads, after some manipulation, to essentially the same prior.

It is a regularizing function with a strength governed by α. Clearly S is more slowly varying than the quadratic regularizer which we looked at in the last section: the contribution to it from a single bin goes as $-f\ln f$ for large f. So it is more likely to be able to pull structure out of the data. In the absence of any data, the most probable image defaults to $\boldsymbol{m}$. Most problems do have a natural baseline that can be used to assign this default. The term $\mu(\boldsymbol{f})$ is separated off from the entropy to avoid including a non-additive term in the expression for S. It can be associated closely with $\mathrm{d}^N\boldsymbol{f}$ as a measure on image space. In this thesis it will be assumed that any posterior distribution of $\boldsymbol{f}$ is sharp enough that the variation of $\mu(\boldsymbol{f})$ across the high probability region of image space is negligible. The expression for $\mathcal{Z}$ arising from (2.17), namely, $(2\pi/\alpha)^{N/2}$, is approximate, valid only when, for each bin, $\alpha m_n \gg 1$. In this case, the probability is well approximated as a Gaussian centred on $\boldsymbol{m}$, a regime whose limitations we investigated in the last section. In other cases, the normalization integral must be evaluated numerically.

With this prior, employed with a Gaussian likelihood, as before, the posterior probability is given by

$$\mathrm{prob(image|data)} \propto \tag{2.22}$$
$$\exp\left(\alpha\sum_n \langle n \mid \left[\mid f\rangle - \mid m\rangle - \mid f\rangle \ln\frac{\langle n \mid f\rangle}{\langle n \mid m\rangle}\right] - \tfrac{1}{2}\sum_p \frac{\left(D_p - \langle p \mid f\rangle\right)^2}{\sigma_p^2}\right).$$

Note that the entropy is applied to a specific representation of the image, that is, the 'bin' representation, in which the image is additive. To find the most probable image $\mid f_0\rangle$, we make small variations in $\mid f\rangle$ and put the coefficient of $\langle n \mid \delta f\rangle$ equal to zero. The process of finding the most probable image is, of course, just the same as the variational problem of maximizing the entropy, subject to a constraint on χ^2. In this old framework, the so-called 'Maximum Entropy Method', the parameter α is related to the Lagrange multiplier used to incorporate the constraint. However the new interpretation (Skilling 1989), in which entropy is a measure of prior probability, gives us a posterior probability distribution of possible images, not just a single best image. From this posterior, we can estimate the likely deviations in the image, as we shall see in Chapter 3.

In this way, we obtain an expression for the most probable image:[10]

$$\langle n \mid f_0\rangle = \langle n \mid m\rangle \exp\left(\alpha^{-1}\sum_p \langle n \mid p\rangle \left(D_p - \langle p \mid f_0\rangle\right)/\sigma_p^2\right). \tag{2.23}$$

Here $\langle n \mid p\rangle$ is the response function for the pth datum in the 'bin' representation. This is not a solution for $\mid f_0\rangle$ but rather a non-linear equation to

[10] For the '$f \sinh^{-1} f$' prior which we shall derive and use in Chapter 3, the equivalent expression has 'sinh' replacing 'exp'.

be solved for it, since $|\, f_0\rangle$ appears on both sides of the equation. Unfortunately, an analytic solution is not available. However, since the likelihood has a non-positive curvature and the prior has a strictly negative curvature everywhere, we can at least be sure that there is a unique solution which is guaranteed to be a maximum.

Although we can't solve the equation in closed form, we can, nevertheless, get some insight into its solution. Remember that image space has two distinct subspaces determined by the kernel, the measured and null spaces ($\mathcal{I} = \mathcal{M} \oplus \mathcal{N}$). We can split $|\, f_0\rangle$ into two parts, its components in each space: $|\, f_0\rangle = |\, f_{\mathcal{M}}\rangle + |\, f_{\mathcal{N}}\rangle$. Our hope is that extrapolation into the null space will have occurred, that is, $|\, f_{\mathcal{N}}\rangle \neq 0$. Now, the $|\, f_0\rangle$ which appears on the RHS of (2.23) is bracketed with each of the response functions $|\, p\rangle$ in turn. But all these are in the measured space and hence we can replace $|\, f_0\rangle$ with $|\, f_{\mathcal{M}}\rangle$ there. In order to obtain an equation exclusively for $|\, f_{\mathcal{M}}\rangle$, it is necessary to project away any of $\mathcal{N}$ in the equation. Remember that the projection operator into measured space is $1_{\mathcal{M}} = \sum_{q\in\mathcal{M}} |\, g_q\rangle\langle g_q\, |$ (see §2.2.3). Applying $1_{\mathcal{M}}|\, n\rangle$ to (2.23) and summing over n gives, on the LHS, $|\, f_{\mathcal{M}}\rangle$. The full equation, for $|\, f_{\mathcal{M}}\rangle$, becomes

$$|\, f_{\mathcal{M}}\rangle = \sum_n 1_{\mathcal{M}}|\, n\rangle\langle n\mid m\rangle \exp\Bigl(\alpha^{-1}\sum_p \langle n\mid p\rangle\bigl(D_p - \langle p\mid f_{\mathcal{M}}\rangle\bigr)/\sigma_p^2\Bigr). \tag{2.24}$$

The solution of this equation would provide the part of the most probable image that is in the measured space, an image described by at most P numbers, the coefficients of a set of vectors spanning $\mathcal{M}$. Once $|\, f_{\mathcal{M}}\rangle$ has been obtained, $|\, f_0\rangle$ is easily derived from (2.23), since, as we said above, the RHS's $|\, f_0\rangle$ can be replaced with the solution for $|\, f_{\mathcal{M}}\rangle$, to give $|\, f_0\rangle$ directly. It is clear that, in general, $|\, f_0\rangle \neq |\, f_{\mathcal{M}}\rangle$ and hence there is some extrapolation into $\mathcal{N}$. This arises, of course, from the exponentiation. For, the argument of the exponential term is a function in $\mathcal{M}$ in a 'bin' representation. However, once it has been exponentiated, there is no reason for the new function to be expandable in the same space. For example, if the argument is an image with a small peak somewhere, the exponentiation might sharpen it up into a image which needs components from $\mathcal{N}$ to describe it. It is interesting to consider what happens if the argument of the exponential is small. The zeroth-order expansion gives $|\, f_0\rangle = |\, m\rangle$: this is the pre-data situation. The first-order expansion, $\exp x \approx 1 + x$, gives the result that is obtained from taking the Gaussian approximation of the prior around the default.

2.3.3 The choice of α

The way to choose α has been the cause of a lot of difficulty. The presence of the regularizer means that the most probable image does not fit the data points exactly. The 'historic' discrepancy method for choosing α (Gull and

Daniell 1978) sprang from this fact. The argument went: on the average, that is over repeated measurements, we expect the deviation of a data point from its noiseless value to be one standard deviation. Therefore, let's insist that the discrepancy between the data points we have measured and the most probable image's predictions of the data averages one standard deviation per point over the single data set we actually have. That is, we use the constraint

$$\sum_{p=1}^{P} \frac{(D_p - \langle p \mid f_0 \rangle)^2}{\sigma_p^2} = P \tag{2.25}$$

to determine α. Since this quantity has the χ^2-distribution of P degrees of freedom, when the errors are independent and Gaussian, it is, itself, often and rather confusingly, called χ^2. Then the constraint to determine α is the very implicit one $\chi^2(\alpha) = P$.

There are many objections to this procedure; some are catalogued by Gull (1989). Briefly, the problem is that this is an orthodox rule. Remember that orthodox statistics relies on the sampling distribution—the probability of the data given the model. But why should we be interested in the long-run behaviour of repeated measurements of the same quantity, when we want to make inferences from the one data set we have? From the pragmatic point of view, the discrepancy method produces images which under-fit the data. The trouble is that, in general, a misfit of P as measured by χ^2 is excessively large. There are well-measured directions in image space which should be fitted rather better than that. Because the most probable image must have this misfit, this image has to change from one that fits all the data points, and since the prior pulls images towards zero, this image will succumb to that pull. As a result the image is biased towards zero and this is transformed into a tendency for data points to be under-fitted. It is really just a more subtle case of the problem we saw with the Gaussian prior in (2.14), where increasing the regularization resulted in a globally reduced image.

In Chapter 3, I suggest and implement a plausible way of overcoming this under-fitting. An alternative, described by Gull (1989), is to treat α as another parameter to be included in the hypothesis space along with the image. This parameter is then marginalized away, or, in practice, fixed at its most probable value. It is the opinion of the present author that more work needs to done on understanding how to estimate α from knowledge of 'macroscopic' properties of the data and/or the image. It is true that α is a dimensional constant and so depends on the units of the image. But from general properties of that image, including the scale on which it is measured, it should be possible to look at a data set and say, "A good α to use here is 0.618 cm." After all, Bayesian reasoning is supposed to model an idealized common sense.

2.3.4 How to compute the most probable image?

The solution of (2.23) has to be obtained numerically, and quite a subtle algorithm is required to obtain it. This thesis presents no new developments in the algorithmics of finding the image that maximizes the posterior probability. Accounts of the Cambridge algorithm are available elsewhere (for example, Skilling and Bryan 1984). The basic idea is to maximize $\alpha S - \frac{1}{2}\chi^2$, iteratively, by making quadratic approximations to this function in a suitably chosen three-dimensional subspace of image space at each iteration. The very non-quadratic form of the function to be maximized renders it essential to employ a non-trivial metric to measure 'distances' between images. This allows intensity to build up in the image in a reasonable time. If α has been estimated beforehand, this is all(!) that has to be done. However, if the discrepancy method is being used to determine α, then α must be juggled at the same time, in order to bring χ^2 to its required value. The reconstructions displayed in this thesis have been obtained by programming this algorithm in `C`. In another method (Gull 1989), the whole optimization process is repeated, with α changed each time, until a criterion determined by the most probable image is satisfied. A large part of the computer's labour is performing the image-to-data transformation and its transpose. Every possible tweak should be made to optimize these matrix operations: 'fast' algorithms, such as the `FFT`, are a *sine qua non*.

It is suggestive that the equation to determine $\mid f_0\rangle$ (2.23) can be reduced to one for $\mid f_{\mathcal{M}}\rangle$ (2.24), from which $\mid f_0\rangle$ is easily derived. It may be possible to exploit this reduction in the scale of the problem, from N equations in N unknowns to P equations in P unknowns, to produce a faster algorithm.

More details of my implementation of the Skilling–Bryan algorithm are given in Appendix A.

2.4 Conclusion

In this chapter, I have discussed in some detail, the Bayesian theory of estimating images specified by a large number of variables from data on some linear transform of them. I have shown how maximizing the likelihood alone fails to produce a good reconstruction because it is unable to say anything at all about a vast number of components of the image. Two different general purpose priors were discussed in detail, and it was demonstrated how the entropic form can produce sharp features in the image, by extrapolating from the data into regions where it has not been measured. The important point, that this kind of free-form image reconstruction can help to suggest more specific models of the image, was made. This is relevant to our analysis of sunspots in Chapter 5.

In the next chapter, I shall put all this to work on a problem in scattering physics, the estimation of nuclear charge distributions.

References

Bretthorst, G. L. (1988). *Bayesian spectral analysis and parameter estimation* (Lecture notes in statistics, Vol. 48). Springer-Verlag, New York.

Daniell, G. J. (1991). Of maps and monkeys: an introduction to the maximum entropy method. In *Maximum entropy in action: a collection of expository essays* (ed. B. Buck and V. A. Macaulay), pp. 1–18. Clarendon Press, Oxford.

Gull, S. F. (1989). Developments in maximum entropy data analysis. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 53–71. Kluwer, Dordrecht.

Gull, S. F. and Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature*, **272**, 686–90.

Gull, S. F. and Skilling, J. (1984). Maximum entropy method in image processing. *IEE Proceedings*, **131F**, 646–59.

Jaynes, E. T. (1986). Monkeys, kangaroos and N. In *Maximum entropy and Bayesian methods in applied statistics: Proceedings of the fourth maximum entropy workshop, University of Calgary, 1984* (ed. J. H. Justice), pp. 26–58. Cambridge University Press.

Jaynes, E. T. (1987). Bayesian spectrum and chirp analysis. In *Maximum-entropy and Bayesian spectral analysis and estimation problems: Proceedings of the third workshop on maximum entropy and Bayesian methods in applied statistics, Wyoming, U.S.A., August 1–4, 1983* (ed. C. R. Smith and G. J. Erickson), pp. 1–37. Reidel, Dordrecht.

Jeffreys, H. (1939). *Theory of probability.* Clarendon Press, Oxford. Later editions: 1948 and 1961.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1986). *Numerical recipes: the art of scientific computing.* Cambridge University Press.

Shore, J. E. and Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, **IT-26**, 26–37.

Sibisi, S. (1990). Quantified MaxEnt: an NMR application. In *Maximum entropy and Bayesian methods, Dartmouth, U.S.A., 1989* (ed. P. F. Fougère), pp. 351–8. Kluwer, Dordrecht.

Skilling, J. (1989). Classic maximum entropy. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 45–52. Kluwer, Dordrecht.

Skilling, J. (1991). Fundamentals of MaxEnt in data analysis. In *Maximum entropy in action: a collection of expository essays* (ed. B. Buck and V. A. Macaulay), pp. 19–40. Clarendon Press, Oxford.

Skilling, J. and Bryan, R. K. (1984). Maximum entropy image reconstruction: general algorithm. *Monthly Notices of the Royal Astronomical Society*, **211**, 111–24.

3
The determination of nuclear charge densities from elastic electron scattering

Schoolmaster: 'Suppose x is the number of sheep in the problem. Pupil: 'But, Sir, suppose x is not the number of sheep.'
J. E. Littlewood: A Mathematician's Miscellany

The determination of nuclear charge densities from data on elastic scattering of electrons from them is a good way to illustrate the methods described in the last chapter. In addition it is a *useful* application, for it reveals an enormous amount that is surprising about just what aspects of these physical distributions are well-determined from currently available data. In the course of the discussion, I shall describe how the reliability of the reconstructed density can be determined.

3.1 The model of the scattering process

For the case of high energy electrons incident upon light nuclei, it is sufficient to describe the scattering by the use of the **first Born Approximation**, a model in which the transition between plane-wave incident and scattered states occurs in a single interaction with the potential generated by the nucleus (Schiff 1949). In this approximation it is not hard to show that the nuclear charge density $\rho(\boldsymbol{r})$ is related to the scattering **form factor** $F(\boldsymbol{q})$ by

$$F(\boldsymbol{q}) = \int \exp(\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r})\rho(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r}, \tag{3.1}$$

in which the variables are $\boldsymbol{r}$, the position in space, and $\boldsymbol{q}$, the momentum transferred in the scattering process. The squared modulus of the form factor is an experimentally accessible quantity: it is the ratio of the differential cross-section for scattering, which is measurable, to that for scattering from an ideal point nucleus of unit charge, which is calculable. From noisy and incomplete measurements of $|F|^2$ at various values of q, we would like to infer ρ. The integral of ρ over a volume gives the amount of enclosed charge and ρ itself is normalized so that this integral is unity. For the case of a spherically symmetric nucleus the angular parts of the above integral can be performed, giving

$$F(q) = \int_0^\infty f(r) j_0(qr)\,\mathrm{d}r, \tag{3.2}$$

with r the radial distance from the centre of the nucleus, q the magnitude of $\boldsymbol{q}$ and $j_0(x) = \sin(x)/x$. The function $f(r)$, which equals $4\pi r^2 \rho(\boldsymbol{r})$, gives, when integrated over a range of r, the amount of charge in a spherical shell: it is the *radial* charge density.[1]

Published data focuses on the differential cross-section or the form factor, one of which is presented in the form of best estimates and expected deviations around those estimates. In an ideal situation, the analyst would have the raw scattering data (from which the published data have been derived) and also a model of the experiment which describes how such data is related to the underlying charge density. This would necessarily be very complicated, and as a result the inferential part of the calculation would be obscured in a lot of very specific technical detail. We avoid such a course and use the already partly processed published data. In any case, the raw data is not readily available.

There exists a lot of prior information about the nucleus suggesting that it is a concentrated distribution, exponentially decreasing outside some small radius, information on, among other things, the finite range of nuclear forces and the perturbation of muonic atomic energy levels. There is, then, a radius R outside of which the amount of charge can be considered altogether negligible and any prior probability of charge can be assigned to be zero.

A small **phase-problem** exists in that only $|F(q)|$ can be derived from the scattering: the sign of $F(q)$ is undetermined. However a clear diffraction structure is visible in the data and there is a prior probability approaching unity (common sense!) that $F(q)$ changes sign as $|F|$ passes through each minimum. This allows us to restore the missing sign, the initial value of which is determined from the normalization requirement $F(0) = 1$. Data points close to diffraction minima, whose signs are doubtful, can be assigned large errors or, equivalently, suppressed. Such non-linear preprocessing of data will reappear when we look at sunspots in Chapter 5.

The range of q values for which $F(q)$ is measured is limited by the experimental arrangement—beam energy and current, for example—but measured q tend to fall quite evenly in a range $[0, q_{\mathrm{max}}]$. At present, for

[1] Often, in the following, the word 'radial' will be omitted. However it is f which is meant and not ρ.

instance, He^4's $q_{\max}$ is $\sim 7\,\mathrm{fm}^{-1}$. Seemingly paradoxical is that such data is both redundant and insufficient. We explore this next.

3.2 The Fourier–Bessel expansion

A first indication of what information is contained in typical electron scattering data can be got by considering the basis-function expansion of f, $f(r) = \sum_n c_n \phi_n(r)$, in terms of the functions

$$\phi_n(r) \propto \begin{cases} r\sin(Q_n r), & \text{if } r < R; \\ 0, & \text{otherwise,} \end{cases} \tag{3.3}$$

where $Q_n = n\pi/R$. This is the **Fourier–Bessel expansion**[2] (Dreher *et al.* 1974). The important thing to observe is the relation between the coefficients c_n of this expansion and the form factor, namely,

$$c_n \propto F(Q_n), \tag{3.4}$$

which is obtained by substituting (3.3) in (3.2). The above relation makes it seem doubtful whether the data can fix more than about the first $Rq_{\max}/\pi$ coefficients, given the maximum momentum transfer achieved in experiment. This conclusion places an upper bound on the number of well measured directions in image space. We can imagine obtaining data points in the region $[0, q_{\max}]$ one by one. Initially they will each provide new information about the charge density, information about the frequency structure near the corresponding momentum transfer. Eventually, however, when their number becomes much larger than the limit mentioned above, the new data are redundant: no fresh information is contained in them.[3] This is the case with the two data sets we shall consider.

However, simultaneously, the data are insufficient, precisely because $F(q)$ is unknown above $q_{\max}$. This corresponds to a lack of measurements of the high frequency structure of the charge density. Any way of inverting such a problem has to make assumptions about this high frequency structure either implicitly, as does the entropic inversion, or explicitly (Dreher *et al.* 1974).

3.3 A prior for a charge density

A Bayesian free-form analysis needs a prior on the charge density, to be plugged into Bayes' theorem. We shall derive one here, for $f(r)$. This

[2] It is called this because it amounts to an expansion of ρ in terms of the functions $j_0(Q_n r)$.

[3] The precise number of points at which this happens depends of course on the quality on the data. The worse it is, the more will be needed.

function is really a mathematical construct, whose only job in life is to be integrated over. That is, it turns out that all observable quantities that might be calculated from it, such as the total charge or the mean-square charge radius, are constructed from integrals linear in f. It is additive: amounts of charge do add up to give the total amount and not combine in some more exotic way, as waves do, for example. Such properties have been raised to the status of entropic axioms (e.g., Skilling 1989, 1991). We demonstrated in Chapter 2 that an entropic prior favours smooth reconstructions and we saw how a simple 'monkey argument' could generate such a prior for a positive distribution. The present distribution is not necessarily positive though. At its most fundamental the nucleus is considered a collection of positively and negatively charged particles, 'quarks', which exist in bound states, the proton and neutron, which have charges one and zero respectively.

Bearing all this in mind, we imagine Chapter 2's team of monkeys, the mechanism of a model satisfying entropic axioms, reappearing equipped now with small elements of both positive and negative charge of the same magnitude ϵ (Buck and Macaulay 1992). These they distribute randomly into a discretized version of the image, of N bins, in such a way that our knowledge of the configuration which results is described by a product of Poisson distributions. We allow that the monkeys might prefer some bins to others and even that this preference might extend to the type of element; as a result, two, possibly different, Poisson parameters ν_i^+ and ν_i^- are assigned to each bin. The probability of a particular configuration is

$$\text{prob}\big(\{n_i^+, n_i^-\}\big) = \prod_{i=1}^{N} \frac{\mathrm{e}^{-\nu_i^+}\, \nu_i^{+\,n_i^+}}{n_i^+!} \frac{\mathrm{e}^{-\nu_i^-}\, \nu_i^{-\,n_i^-}}{n_i^-!}, \tag{3.5}$$

with n_i^+ (n_i^-) the number of positive (negative) elements in the ith bin. By construction, inference about one bin is independent of any other, and so we can focus attention on one, the ith, say, and, for the moment, suppress its label. We should like to calculate the probability of the total charge $\epsilon(n^+ - n^-)$ in that bin. Define, for convenience, the new variables

$$\begin{aligned} n &= \tfrac{1}{2}\big(n^+ + n^-\big) \\ \text{and} \qquad q &= \tfrac{1}{2}\big(n^+ - n^-\big), \end{aligned} \tag{3.6}$$

each of which can take on integer or half-integer values, restricted only by the requirement $|q| \leq n$. The aforementioned amount of charge is then $2\epsilon q$. The probability of the new variables satisfies

$$\text{prob}\big(n, q\big) = \mathrm{e}^{-(\nu^+ + \nu^-)} \gamma^q \frac{\beta^{2n}}{(n+q)!\,(n-q)!}, \tag{3.7}$$

where $\beta^2 = \nu^+\nu^-$ and $\gamma = \nu^+/\nu^-$. Now, to obtain the probability of q, it is necessary to marginalize away n, that is,

$$\mathrm{prob}(q) = \sum_{n\geq|q|} \mathrm{prob}(n,q). \tag{3.8}$$

This sum can be performed exactly (Lebedev 1972, p. 108) and yields

$$\mathrm{prob}(q) = \mathrm{e}^{-(\nu^+ + \nu^-)}\gamma^q I_{2q}(2\beta). \tag{3.9}$$

Here $I_n(x)$ is the modified Bessel function of order n and argument x. As usual we envision the element of charge becoming very small, so that q can be considered very large. Hence, using a result on the approximate form of Bessel functions of large order,[4] we gain the important result

$$\begin{aligned}\mathrm{prob}(q) &\sim \frac{1}{\sqrt{2\pi}}\big((2q)^2 + (2\beta)^2\big)^{-\frac{1}{4}} \\ &\times \exp\left(\sqrt{(2q)^2+(2\beta)^2} - \beta\big(\gamma^{\frac{1}{2}} + \gamma^{-\frac{1}{2}}\big) + q\ln\gamma - 2q\sinh^{-1}\frac{q}{\beta}\right).\end{aligned} \tag{3.10}$$

In practice, it is usually more useful to consider a continuum limit, where the discrete q is replaced by a continuous f, the total charge in the bin; it is the probability density of this variable which is needed. The identification $\sum_q \to \int \mathrm{d}f/\epsilon$ is made, with the measure ϵ^{-1} arising from the spacing of the underlying discrete values of f $(= 2\epsilon q)$, and we find

$$\mathrm{prob}\big([f, f+\mathrm{d}f]\big) \simeq \sqrt{\frac{\alpha}{2\pi}}\exp\big(\alpha s(f,m,\gamma)\big)\,\mu(f)\,\mathrm{d}f. \tag{3.11}$$

Definition of terms and some comment is called for. We have employed $\alpha = \epsilon^{-1}$ and $m = 2\beta\epsilon$. The function $s(f,m,\gamma)$, defined by

$$s = \sqrt{f^2+m^2} - \tfrac{1}{2}m\big(\gamma^{\frac{1}{2}} + \gamma^{-\frac{1}{2}}\big) + \tfrac{1}{2}f\ln\gamma - f\sinh^{-1}\frac{f}{m}, \tag{3.12}$$

is the contribution from a single bin to a new configurational entropy $S = \sum_i s(f_i, m_i, \gamma_i)$, while the function μ is defined by

$$\mu(f) = (f^2+m^2)^{-\frac{1}{4}}, \tag{3.13}$$

and can be associated closely with $\mathrm{d}f$ as a measure appropriate to a curved image space.[5] The normalization of (3.11) is approximate, correct only

[4] Meissel's first extension of Carlini's formula (Watson 1944).

[5] This interpretation proves useful when it comes to constructing efficient entropy maximizing algorithms.

in the limit that the distribution is effectively Gaussian, that is, when $\alpha m \gg 1$. In general the normalization factor Z (which replaces $\sqrt{2\pi/\alpha}$) can be written as $\sqrt{m}\, z(\alpha m)$, where the function $z(x) \underset{x\to\infty}{\sim} \sqrt{2\pi/x}$.

The constants m and γ deserve some explanation. Both are parameters which must either be assigned *a priori* or be treated as additional quantities to be inferred or marginalized away. They allow for extra flexibility in the model of our state of knowledge of the nucleus or whatever image the prior is being used to infer. Whereas the traditional entropic prior of a positive distribution has only one such adjustable parameter,[6] namely, the baseline parameter m, the 'sinh' form just derived has two, which perform similar functions. The new m is of the same dimensions as the image f for which it sets the scale: the larger it is, the larger the image must be before the non-linearity of s is felt. It is directly related to the prior expected width of the distribution about its mean. The γ parameter allows the most probable image $\hat{f}$ to be shifted away from zero, for $\hat{f} = m\sinh(\frac{1}{2}\ln\gamma)$. In Gaussian approximation this is also the expected image $\langle f\rangle$. Indeed we can use this result to eliminate γ in favour of $\langle f\rangle$ to give the pleasingly symmetrical result

$$s(f) = \sqrt{f^2+m^2} - \sqrt{\langle f\rangle^2+m^2} + f\Big(\sinh^{-1}\frac{\langle f\rangle}{m} - \sinh^{-1}\frac{f}{m}\Big). \qquad (3.14)$$

In the sequel, $\langle f\rangle$ will be assigned to be zero so as to preserve positive/negative symmetry.

The constant α determines how much smoothing the prior is able to do. When it is very small, the prior distribution is broad and hence effectively constant over the set of plausible images. Thus, very little smoothing results. At the other extreme, when α is large, the prior is very sharply peaked around zero. The posterior is pulled towards this value and a smooth image, biased towards zero, results. Between these two regimes lies a region of useful values, where effective smoothing is achieved with tolerable bias.

To summarize, we put all N bins back together and show the prior density for the full image $\boldsymbol{f} \equiv (f_1, \ldots, f_N)^{\mathrm{T}}$:

$$\mathrm{prob}(\boldsymbol{f}|\alpha, \boldsymbol{m}) = \frac{1}{\mathcal{Z}(\alpha, \boldsymbol{m})}\exp\big(\alpha S(\boldsymbol{f}, \boldsymbol{m})\big)\mu(\boldsymbol{f}, \boldsymbol{m}), \qquad (3.15)$$

where

$$S(\boldsymbol{f}, \boldsymbol{m}) \;=\; \sum_i \left(\sqrt{f_i^2+m_i^2} - m_i - f_i\sinh^{-1}\frac{f_i}{m_i}\right), \qquad (3.16)$$

[6] apart from the regularization strength α, which governs the relative weighting given to the data vs prior smoothing.

$$\mu(\boldsymbol{f}, \boldsymbol{m}) = \prod_i \left(f_i^2 + m_i^2\right)^{-\frac{1}{4}} \qquad \text{and} \tag{3.17}$$

$$\mathcal{Z}(\alpha, \boldsymbol{m}) = \int \exp\left(\alpha S(\boldsymbol{f}, \boldsymbol{m})\right) \mu(\boldsymbol{f}, \boldsymbol{m})\, \mathrm{d}^N \boldsymbol{f}. \tag{3.18}$$

3.4 Inferring charge densities

For this application, the natural identification of each f_i is with the amount of charge in a spherical shell, which equals $\int_{i\text{th shell}} f(r)\,\mathrm{d}r$ in the notation of §3.1. We will always take enough bins (of width Δ_i) so that this expression can be well approximated by $f(r_i)\Delta_i$, with $r_i \in i$th shell. In addition, for shells of equal thickness, the common Δ equals R/N and r_i is taken as $(i + \frac{1}{2})\Delta$.[7] Formally, at least, we could replace (3.15)–(3.18) by relations applicable to *functions* $f(x)$, so that the entropy would become a function*al* of $f(x)$ and probabilities functional integrals over probability functionals. However to evaluate such integrals it would be necessary to use (3.15)–(3.18) to approach the continuum as a limit, a limit which is not of use in application. As a result, this step is not taken.

The assignment of the likelihood proceeds straightforwardly. Having available best estimates and expected deviations of each datum and an additive noise model $D_p = F_p + n_p$, with D_p a datum, n_p the noise in that datum and $F_p = F(q_p)$ the true form factor, we use Jaynes' principle, as outlined in Chapter 1, to derive the Gaussian distribution, $\text{prob(data|image)} = \text{prob}(\{D_{p=1\to P}\}|\{f_{i=1\to N}\}) \propto \exp(L)$, where $L = -\frac{1}{2}\sum_p (D_p - \hat{O}_p \boldsymbol{f})^2 \sigma_p^{-2}$. Here $\hat{O}_p$ is the image-to-data transformation for the pth datum, $\sum_r \cdots j_0(q_p r)$ (the variable r of (3.2) has been discretized), and σ_p^2 is the estimated squared error in D_p; the number of data is P and the discretized image $\{f_i\}$ has N bins; the quantity L is called the *log-likelihood*. The measure of misfit χ^2 equals $-2L$. Although a Poisson distribution might be more appropriate for the description of what is essentially a counting process, the published cross-sections are so processed from the underlying raw counts, that falling back on the more primitive maximum entropy probability short-circuits the tedious business of calculating the derived distribution. This outweighs any small loss of resolution that might be incurred by using this more conservative Gaussian distribution.

The choice of the scale-setting parameters $\{m_i\}$ that appear in our entropy formula is considered next. Again a very conservative assignment, based on very general arguments is most satisfactory, avoiding as it does the charge of arbitrariness. A natural scale is set by the requirement that the total charge in $[0, R]$ must add up to unity. In order to convert this into an assignment of m_i in each bin, we suppose that the charge is distributed

[7] In this case, Δ could be factored out of the definition of f_i, leaving the entropy a function of a charge *density*.

as evenly as possible over the *a priori* feasible sphere of radius R, producing the uniform charge density. Then the amount of charge in the ith spherical shell, a zeroth-order estimate of f_i, is proportional to r_i^2, because of the volume factor. This seems a reasonable first guess at the scale of the charge in the ith bin. So, supplying the constant explicitly, we arrive at the fully-determined scale function

$$m_i = \begin{cases} \dfrac{3\Delta}{4\pi R^3} r_i^2, & r_i \leq R; \\ 0, & r_i > R. \end{cases} \tag{3.19}$$

(cf. Daniell and Potton 1989). Despite its discontinuous behaviour at the edge of the nucleus, it serves well.

As we briefly mentioned in Chapter 2, the most difficult parameter to assign is the prior–likelihood relative weighting factor, α. We mentioned there the 'historic' discrepancy method as a flawed way of determining it. Remember that it leads to a general under-fitting of the data. We are reluctant to increase the hypothesis space to include α as another parameter to estimate from the data, as other have done (Gull 1989); the computational burden becomes very much larger, and even then some approximations are necessary.[8]

We propose modifying the 'historic' method in a plausible way to overcome the under-fitting problem. There are directions in image space that are measured very well by a particular data set, but the constraint $\chi^2 = P$ insists that every direction, on average, pays its unit of misfit. Because the prior is peaked around zero, this misfit will be determined in such a way as to pull the reconstruction towards zero. However if we knew how many directions were well-measured, we could exempt them from the misfit tax and hence reduce the systematic under-fitting. Then the new constraint $\chi^2 = P - W$ could be used, where W is the number of 'good' measurements. Fortunately, it is not too hard to estimate W for the two data sets I shall be considering. It was already mentioned in §3.2 that the number of independent data is limited by the maximum momentum transfer attained in each experiment and an estimate of that number was obtained on the assumption that all the directions that could be determined by the data beneath $q_{\max}$ were well measured. The value for W determined in this way agrees well with another natural measure of the number of good directions. This is obtained by diagonalizing the kernel in the Fourier–Bessel basis. When the eigenvalues are put in decreasing order of size, it is clear how many correspond to well measured directions because after that

[8] For example, it is not practical to marginalize α away from prob$(\alpha, f|\text{data})$ although this is the thoroughgoing Bayesian technique for dealing with a nuisance parameter. Instead the α which maximizes prob$(\alpha|\text{data})$, α_0, say, is used in prob$(f|\alpha_0,\ \text{data})$, which in turn can be used to determine the most probable image. This is only an approximate scheme though and it relies on the probability of α being very sharply peaked.

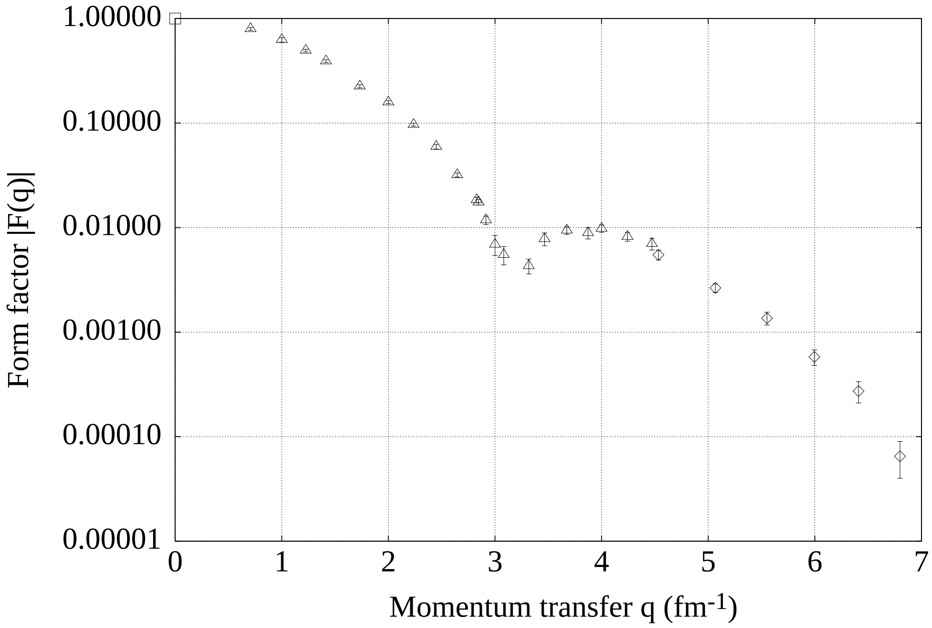


Fig. 3.1. The He^4 data set ($\triangle$ Frosch *et al.* 1967; $\diamond$ Arnold *et al.* 1978; $\square$ Normalization).

number they start decreasing much more rapidly: this is the 'eigenvalue cliff' (Macaulay and Buck 1990). A full Bayesian treatment allows W to be estimated in yet another way (Gull 1989).

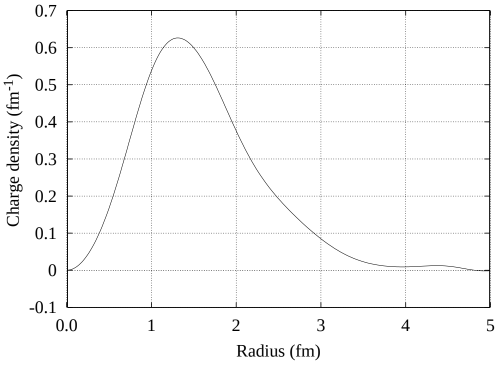
In the next section, I shall summarize the results of an analysis of the data for He^4 and C^{12}. In Appendix A, I discuss how these results were computed.

3.5 Two examples: He^4 and C^{12}

3.5.1 He^4

Figure 3.1 shows experimental data on the form factor of He^4 up to $q \sim 7\,\mathrm{fm}^{-1}$ (Frosch *et al.* 1967; Arnold *et al.* 1978). Measurements close to the diffraction minimum have been omitted: their phase is uncertain and the corrections to the Born approximation there are potentially large. The extra point $F(0) = 1$ has been added with a small error (± 0.01) to enforce normalization. In Figure 3.2, the most probable charge density is shown in two forms, the most probable $f(r)$ and the corresponding $\rho(r) = f(r)/(4\pi r^2)$. The parameters are as follows: R is taken as $5\,\mathrm{fm}$, $q_{\mathrm{max}} \approx 7\,\mathrm{fm}^{-1}$ and P equals 28. Thus $W \approx R q_{\mathrm{max}}/\pi \approx 11$ good independent data. The number of large eigenvalues of the kernel in the Fourier–Bessel basis is nearer 12, which suggests that some extrapolation of the data beyond q_{max} is possible. Thus we use the criterion $\chi^2 = 28 - 12 = 16$ to fix on a reasonable α.

It is of considerable interest that this charge density is almost every-



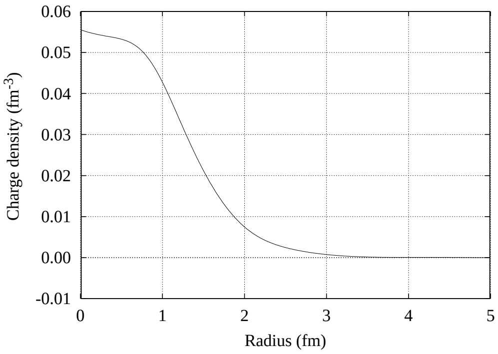


Fig. 3.2. The most probable charge density of He^4. Two forms are shown: on the left $f(r)$ and on the right $\rho(r) = f(r)/(4\pi r^2)$. Notice that, although the prior has no bias towards positive or negative values, there is almost no negative density in either figure.

where positive. We hedged our bets about this by using a prior which favoured neither positive nor negative charge, and yet the data was strong enough to predict a positive density. That this would occur could not have been anticipated just from looking at the data. In §3.6, when we look at error-bars, we shall see that the negative excursions are probably not significant. We shall also see how this model interpolates within the given data.

It is worth saying that this image of the He^4 charge density, and indeed the ones of C^{12} to follow, are very similar in form to reconstructions made using other techniques (e.g. Frosch 1967). But whereas those analyses largely have employed explicit models of the nuclear shape, with a handful of parameters, there is no such model here; the data are left to speak for themselves, restrained only by a prior which favours a smooth density and which can be justified on very general grounds.

3.5.2 C^{12}

Figure 3.3 shows experimental data on the form factor of C^{12} up to $q \sim 3.5\,\mathrm{fm}^{-1}$ (Jansen *et al.* 1972; Sick and McCarthy 1970). This data is more problematical than that for He^4: there is evidence for a systematic incompatibility between the data measured at different energies by Sick and McCarthy. It shows up in the region where the sets overlap: a line drawn through the data points zigzags from side to side, that is, one set is consistently larger than the other. The effect is larger than the error-bars. This I have taken as a sign that the errors quoted are too optimistic and as a result they have been doubled throughout. In addition, in the region where the sets overlap, only the lower energy data has been retained. Again points have been omitted from the diffraction minimum. Figure 3.4 shows the most probable densities $f(r)$ and $\rho(r)$. The relevant parameters are: $R = 6\,\mathrm{fm}$, $q_{\mathrm{max}} \approx 3.6\,\mathrm{fm}^{-1}$ and $P = 42$. The number of good directions

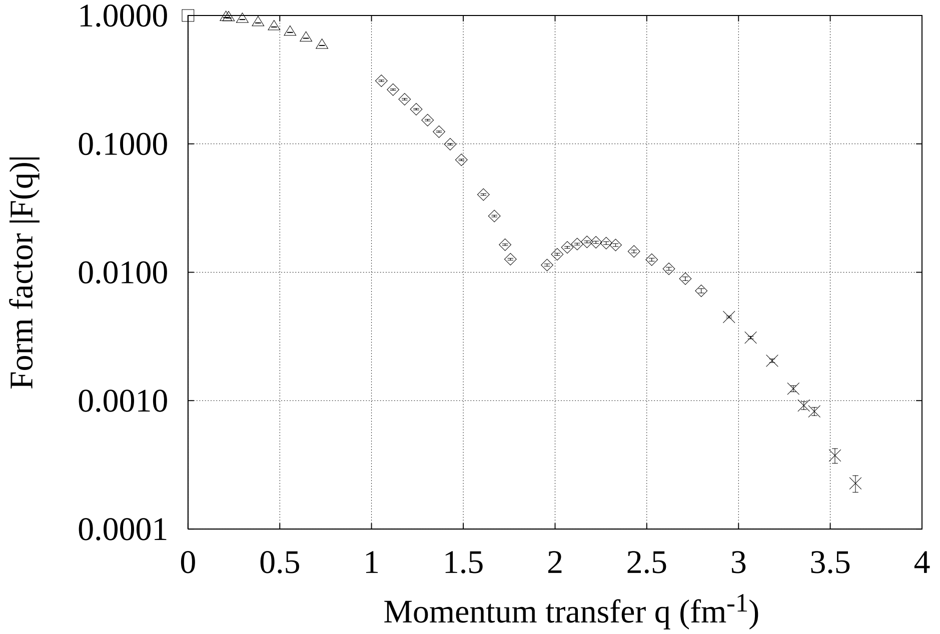


Fig. 3.3. The C^{12} data set ($\triangle$ Jansen *et al.* 1972; $\diamondsuit$ 374.5 MeV, $\times$ 747.4 MeV, Sick and McCarthy 1970; $\square$ Normalization).

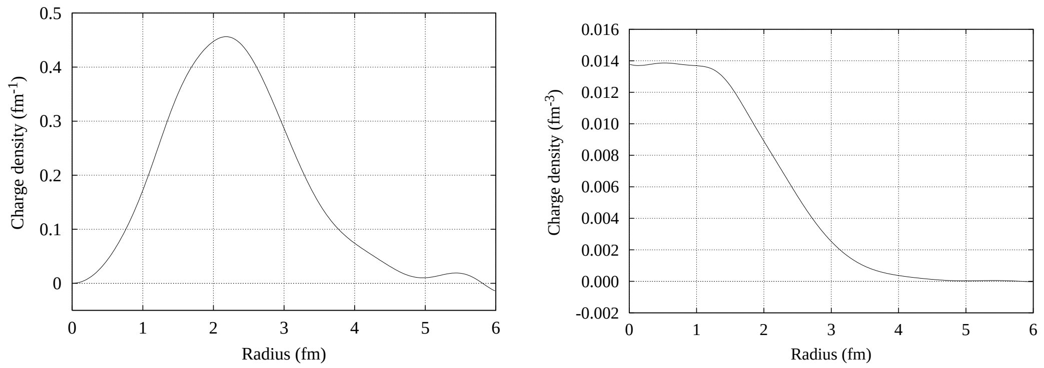


Fig. 3.4. The most probable charge densities $f(r)$ (left) and $\rho(r)$ (right) for the nucleus C^{12}.

is ∼7, giving a target χ^2, W, of approximately 35.

The behaviour of f above 4 fm is interesting. In the analysis of Sick and McCarthy (1970), which was a fit of an explicit model of the nuclear shape to the data, a Gaussian was introduced into the model at the surface of the nucleus in order to improve the fit to the data. However, our error analysis in the next section suggests that this feature is probably not significant. This shows two things. Firstly, non-Bayesian parameter estimation techniques, such as least-squares, provide no way of ruling out overcomplicated models which end up fitting noise. That is, they do not contain an internal Ockham's Razor.[9] Secondly, it makes plain how risky it is to rely on one 'best' image, unqualified by error estimates. We remedy that next.

3.6 Estimation of expected deviations

3.6.1 The theory

We now look at the question of the reliability of our reconstructed charge densities. The traditional way to do this is to quote expected deviations around the best image—error-bars, or, for a continuous image, error-bands. In the early years of the development of entropic image restoration, when the 'Maximum Entropy Method' was a variational principle by which a 'best' image was chosen (Gull and Daniell 1978), there was no rationale for producing error estimates. However, now that the entropy function is seen as being related to a rather useful prior probability density, there is no problem in determining expected errors, at least in principle. The situation is quite reversed: there is an *embarras de choix*. For we can construct the entire posterior probability distribution, a distribution of many variables, which Skilling (1990) has rather aptly called a probability **bubble**, full of information on subtle correlations and the like, which is the complete inference about the image in the light of the data. However it is not feasible or useful to record this gigantic beast.[10] We must settle for a more manageable summary of the information in this distribution. Even the humble covariance matrix, the elements of which are $\big\langle (f_i - \langle f_i \rangle) \times (f_j - \langle f_j \rangle) \big\rangle$ and which measure the correlations in the probabilities of each pair of variables, is too cumbrous, having order N^2 different entries. The use of error-bars, which are the square-roots of the diagonal elements of this matrix, is a feasible compromise. However they have to be viewed in an entirely fresh way because of how they behave when the image is discretized

[9] This is to be contrasted with the Bayesian approach to the same problem (MacKay 1992).

[10] A novel technique which gives a singular insight into this function has been developed by Skilling *et al.* (1991). It consists of displaying a film sequence of samples from the posterior, which makes correlations in the inference very apparent. Now that journals are beginning to accept video-recordings as a part of papers (Lindley 1992), this clever, but computationally expensive, approach might yet be applied to real problems!

at different coarsenesses. For, their sizes depend strongly on how finely the image is sampled: they increase without limit as the number of bins is increased, that is, as the continuum (error-*band*) limit is approached. That is worth repeating: error-bands diverge for this model.

This is not a disaster. On the contrary, it is exactly in line with (educated) common sense. (Often it takes the execution of a Bayesian analysis to reveal what should have been apparent all along.) As the bin width narrows, the hypothesis space is enlarged to include propositions about progressively finer and finer details of the distribution of charge. But does it not correspond nicely with intuition that, *a priori*, the microstructure of an image is more uncertain than its macrostructure? In fact this behaviour persists in the posterior; even the data cannot help here. For there is always some minimum wavelength, that is, maximum q, at which data is available. Even though some extrapolation to smaller scale (shorter wavelength) structure might be possible, the data will not be able to say anything about structure on very much smaller scales. If the hypothesis space contains propositions about such small features, we can expect large error-bars. In the limit of infinitely intricate detail, we can make no reliable prediction at all.

Again a statistical mechanical analogy is suggested. A datum on the mean energy of a gas at equilibrium in given external conditions is often sufficient for us confidently to predict other *macroscopic* properties of the system, for example, its pressure. But if we were asked to predict the pressure at an instant on progressively smaller and smaller sections of the walls of the containing vessel, we would become more and more uncertain about the result, that is, the expected deviations would increase.

More informative priors would be necessary to do better than this. Hypotheses on the asymptotic behaviour of the form factor have been used to curb the total ignorance about fine scale structure, but hypotheses they really are, designed specifically to control diverging error-bands (Dreher *et al.* 1974 and, in a Bayesian context, Macaulay and Buck 1990). Such methods' error-bars are necessarily conditioned on the hypotheses employed, and lack the intuitive generality of the MaxEnt results, which are based on very few assumptions about the image. In the two cited papers, the hypothesis is that the second derivative of the density has a certain sort of smoothness.[11] This places a constraint on the form of $F(q)$ as $q \to \infty$, and hence on the coefficients in the Fourier–Bessel expansion. Additional assumptions, the validity of which is not entirely clear, allow for the construction of a probability for the charge density which leads to finite error-bands.

Let us see how the dependence of error-bars on scale arises in the MaxEnt model. As I have said, the behaviour can be seen already in the prior. By the *a priori* assumption of no interbin correlation, the prior covariance

[11] Specifically, *limited total fluctuation.*

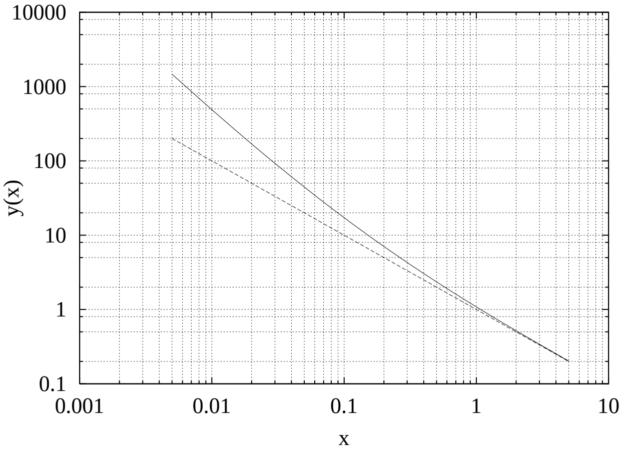


Fig. 3.5. A log–log plot of the functions $y(x)$ (solid) and x^{-1} (dashed).

matrix for the charge is diagonal and its (i,i)th element is written most generally as $m_i^2 y(\alpha m_i)$, where y is some function: for example, it is not hard to show that $y(x) \underset{x\to\infty}{\sim} x^{-1}$. Indeed the monotonic increase of this function as $x \to 0^+$ is even faster than x^{-1}, so that, as $x \to 0$, $y(x) \to \infty$: see Figure 3.5. The squared-error on the charge *density* in the ith bin is

$$\left\langle \big(f(r_i) - \langle f(r_i)\rangle\big)^2 \right\rangle = \left\langle f^2(r_i)\right\rangle = \left(\frac{m_i}{\Delta}\right)^2 y(\alpha m_i). \tag{3.20}$$

As the bin size Δ is decreased, how do the parameters vary? As long as Δ is small enough that no plausible structure in the image is under-sampled, the log-prior/log-likelihood weighting factor α should be unchanged. For, both the entropy and the transforms of the image that appear in χ^2 are in essence integrals over the variable r which are being approximated by sums. As the number of bins increases, all that happens is that these integrals are better and better approximated. Thus the entropy and χ^2 are essentially unchanged, and hence so is their relative weight. However m_i is an integral of the scale-setting function over a bin. As long as the image does not change much over that bin, we would expect the integral to decrease in proportion to the size of the bin. So each m is roughly proportional to Δ. But this is all we need. For as $N \to \infty$ and $\Delta \to 0$, then, in each bin, $\alpha m \to 0$, while $(m/\Delta)^2$ is constant, so that the squared-error on the charge density in a bin (3.20) goes to infinity.

Bearing all this mind, let us proceed by calculating the posterior covariance matrix. Its (i,j)th element C_{ij} is determined from

$$\begin{aligned} C_{ij} &\equiv \big\langle (f_i - \langle f_i\rangle)(f_j - \langle f_j\rangle)\big\rangle \qquad (3.21)\\ &= \frac{\int_{\mathcal{I}} (f_i - \langle f_i\rangle)(f_j - \langle f_j\rangle)\exp\big(\alpha S(\boldsymbol{f}) + L(\boldsymbol{f})\big)\,\mu(\boldsymbol{f})\,\mathrm{d}^N\boldsymbol{f}}{\int_{\mathcal{I}} \exp\big(\alpha S(\boldsymbol{f}) + L(\boldsymbol{f})\big)\,\mu(\boldsymbol{f})\,\mathrm{d}^N\boldsymbol{f}}, \end{aligned}$$

where, as a reminder, $\mathcal{I}$ signifies all possible images, f_i is the image strength in the ith bin, α is the regularizing constant, S and μ are, respectively, the entropy and the measure derived in §3.3, L is the log-likelihood and $\langle\cdots\rangle$ denotes an average over the probability of the image. These are two integrals which cannot be calculated exactly. Instead we assume that their common exponential term is sufficiently sharply peaked about its (single) maximum that the method of steepest descents provides good values for them. The effect of this approximation is that the argument of the exponential is expanded to second-order in deviations away from the most probable image $\boldsymbol{f}_0$, and the comparatively slowly varying $\mu(\boldsymbol{f})$ takes on the constant value $\mu(\boldsymbol{f}_0)$, since it is only when $\boldsymbol{f} \approx \boldsymbol{f}_0$ that it is not damped away by the exponential. Now

$$\begin{aligned}\alpha S + L &= \text{constant} + \delta\boldsymbol{f}^{\mathrm{T}}\left[\alpha\nabla S + \nabla L\right]_{\boldsymbol{f}_0} \\ &\quad - \tfrac{1}{2}\delta\boldsymbol{f}^{\mathrm{T}}\mathbf{C}^{-1}\delta\boldsymbol{f} + O\big((\delta\boldsymbol{f})^2\big),\end{aligned} \tag{3.22}$$

where

$$\mathbf{C}^{-1} = \mathbf{g}\big(\alpha\mathbf{I} - \mathbf{g}^{-1}\left[\nabla\nabla L\right]_{\boldsymbol{f}_0}\mathbf{g}^{-1}\big)\mathbf{g} \tag{3.23}$$

with $g_{ij} = \delta_{ij}\sqrt{-\partial^2 S/\partial f_i^2}\big|_{\boldsymbol{f}_0}$ (no summation). Here $\delta\boldsymbol{f} \equiv \boldsymbol{f} - \boldsymbol{f}_0$, the operator $\nabla = \big(\partial/\partial f_1, \ldots, \partial/\partial f_N\big)^{\mathrm{T}}$ and $\nabla\nabla$ is a tensor operator, the (i,j)th element of which is $\partial^2/\partial f_i \partial f_j$. Because there is no correlation between bins in the prior, $\nabla\nabla S$ is diagonal which explains why $\mathbf{g}$ also is diagonal and hence why it never appears transposed. Using the functional form of S (3.16), we see immediately that $\mathbf{g} = \mathrm{diag}\big((f_i^2 + m_i^2)^{-\frac{1}{4}}\big)$. The constant in (3.22) can be ignored and the coefficient of the term linear in $\delta\boldsymbol{f}$ vanishes at $\boldsymbol{f}_0$ since it is just the condition that $\boldsymbol{f}_0$ be the most probable image. The two integrals of (3.21) then have the approximate form

$$\mu(\boldsymbol{f}_0)\int_{\mathcal{I}} \cdots \exp\big(-\tfrac{1}{2}(\boldsymbol{f} - \boldsymbol{f}_0)^{\mathrm{T}}\mathbf{C}^{-1}(\boldsymbol{f} - \boldsymbol{f}_0)\big)\,\mathrm{d}^N\boldsymbol{f}. \tag{3.24}$$

The exponential is a multivariate Gaussian (Martin 1971). It is not too hard to show that, for this distribution, the mean image $\overline{\boldsymbol{f}}$ equals $\boldsymbol{f}_0$, and the covariance matrix of the variables is $\mathbf{C}$.[12] Its inverse $\mathbf{C}^{-1}$ is rather interesting. It is manifestly symmetric and, in fact, also positive-definite, since

[12] Suppose $p(f) \propto \exp(-\frac{1}{2}(f-f_0)^{\mathrm{T}}\mathsf{C}^{-1}(f-f_0))$. Assume C is symmetric and positive-definite. Therefore it can be diagonalized: $\mathsf{C} = \mathsf{U}\Lambda\mathsf{U}^{\mathrm{T}}$ with $\mathsf{U}^{\mathrm{T}}\mathsf{U} = \mathsf{U}\mathsf{U}^{\mathrm{T}} = \mathsf{I}$, so that $\mathsf{C}^{-1} = \mathsf{U}\Lambda^{-1}\mathsf{U}^{\mathrm{T}}$. Change variables: $y = \mathsf{U}^{\mathrm{T}}(f - f_0)$. Then, since $\det\mathsf{U} = 1$, $p(y) \propto \exp(-\frac{1}{2}y^{\mathrm{T}}\Lambda^{-1}y)$. But Λ is the diagonal matrix of eigenvalues, so that this distribution factorizes into separate distributions for each y_i with $\langle y_i\rangle = 0$ and $\mathrm{cov}(y_i, y_j) \equiv \langle(y_i - \langle y_i\rangle)(y_j - \langle y_j\rangle)\rangle = \langle y_i y_j\rangle = \delta_{ij}\Lambda_i$ (no summation). Since $f = \mathsf{U}y + f_0$, $\langle f\rangle = \mathsf{U}\langle y\rangle + f_0 = f_0$. Then $\mathrm{cov}(f_i, f_j) = \langle U_{ik}y_k U_{jl}y_l\rangle = U_{ik}U_{jl}\langle y_k y_l\rangle = U_{ik}U_{jl}\,\mathrm{cov}(y_k, y_l) = U_{ik}\delta_{kl}\Lambda_k U_{jl} = \mathsf{U}\Lambda\mathsf{U}^{\mathrm{T}}\big|_{ij} = C_{ij}$, as claimed.

$\alpha\mathbf{I}$ and $\mathbf{g}$ are positive-definite while $-\nabla\nabla L$ is positive-semi-definite. For the Gaussian likelihood, this Hessian $-\nabla\nabla L$ is just the *kernel* of Chapter 2.

In approximating the posterior density as a Gaussian around the most probable image, we have ended up with a distribution strikingly similar to that which emerged from our consideration of a Gaussian prior in Chapter 2. Of course the most probable image is in general quite different, but the structure of the inverse-covariance is familiar: a weighted identity matrix added to the kernel.[13] In order to obtain error-bars, we must invert $\mathbf{C}^{-1}$. This is a large beast, in fact, $N \times N$. If this matrix were not to have some special properties, we would be stymied; the computational price would be prohibitive. Fortunately, being built around the kernel, it *is* rather special. We showed that the kernel has at most P eigenvectors of non-zero eigenvalue. All remaining eigenvectors are not measured by the data—they have zero eigenvalue—and they are regularized by the term $\alpha\mathbf{I}$. This provides a further insight into α: it is the quantity which determines how sharp the inference about unmeasured directions will be. The presence of $\mathbf{g}$ does not radically alter this result; all that happens is that the measured space is slightly altered. Suppose we could diagonalize $-\mathbf{g}^{-1}\nabla\nabla L\mathbf{g}^{-1}$ so that it could be written as $\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^{\mathrm{T}}$, with only up to the first P elements of the diagonal of $\mathbf{\Lambda}^2$, the eigenvalue matrix, being non-zero. (The matrix $\mathbf{U}$ has the eigenvectors as columns.) This is a large $N \times N$ problem but we shall only need to do part of it. With

$$\mathbf{C}^{-1} = \mathbf{g}(\alpha\mathbf{I} + \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^{\mathrm{T}})\mathbf{g}, \tag{3.25}$$

it is not hard to show[14] that

$$\mathbf{C} = \alpha^{-1}\mathbf{g}^{-1}\left(\mathbf{I} - \mathbf{U}\mathbf{\Lambda}^2(\alpha\mathbf{I} + \mathbf{\Lambda}^2)^{-1}\mathbf{U}^{\mathrm{T}}\right)\mathbf{g}^{-1}. \tag{3.26}$$

But this is very easy to compute because the only parts of $\mathbf{U}$ which are needed are the eigenvectors corresponding to non-zero Λ_{ii}. (The other eigenvectors get multiplied by their eigenvalue and hence make no contribution.) There are only up to P of these, which can be found by solving a much smaller $P \times P$ problem. This is a result of the singular value decomposition (see Chapter 2). Let $\mathbf{O}$ be the $P \times N$ matrix that performs the image-to-data transformation, and $\mathbf{\Sigma}$ the $P \times P$ diagonal covariance matrix of the

---

[13] The matrix g intrudes though. It can be magicked away, however, by reinterpreting image-space to be curved, with a metric $-\nabla\nabla S = \mathsf{g}^2$. We shall not use this device here.

[14] For example, using the fact that $\mathsf{U}\mathsf{U}^{\mathrm{T}} = 1$, we can write $\mathsf{C}^{-1} = \mathsf{g}\mathsf{U}(\alpha\mathsf{I} + \Lambda^2)\mathsf{U}^{\mathrm{T}}\mathsf{g}$. Then

$$\begin{aligned} \mathsf{C} &= \mathsf{g}^{-1}\mathsf{U}(\alpha\mathsf{I} + \Lambda^2)^{-1}\mathsf{U}^{\mathrm{T}}\mathsf{g}^{-1} \\ &= \alpha^{-1}\mathsf{g}^{-1}\mathsf{U}(\mathsf{I} - \Lambda^2(\alpha\mathsf{I} + \Lambda^2)^{-1})\mathsf{U}^{\mathrm{T}}\mathsf{g}^{-1} \\ &= \alpha^{-1}\mathsf{g}^{-1}(\mathsf{I} - \mathsf{U}\Lambda^2(\alpha\mathsf{I} + \Lambda^2)^{-1}\mathsf{U}^{\mathrm{T}})\mathsf{g}^{-1}, \end{aligned}$$

as required.

data, $\mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_P^{-2})$. Then the modified Hessian $-\mathbf{g}^{-1}\nabla\nabla L\mathbf{g}^{-1}$ equals $\mathbf{g}^{-1}\mathbf{O}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{O}\mathbf{g}^{-1}$. Define $\mathbf{M} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{O}\mathbf{g}^{-1}$. It is not hard to show that the $P \times P$ matrix $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ has the same non-zero eigenvalues as $-\mathbf{g}^{-1}\nabla\nabla L\mathbf{g}^{-1} = \mathbf{M}^{\mathrm{T}}\mathbf{M}$. Let us collect together the P-dimensional eigenvectors corresponding to these eigenvalues in the $P \times P$ matrix $\mathbf{V}$. Then the singular value decomposition gives $\mathbf{M} = \mathbf{V}\boldsymbol{\Lambda}_{\mathrm{P}}\mathbf{U}_{\mathrm{P}}^{\mathrm{T}}$, where the subscript P on $\boldsymbol{\Lambda}$ and $\mathbf{U}$ indicates that, respectively, zero eigenvalues and their eigenvectors have been removed, leaving matrices of sizes $P \times P$ and $N \times P$. This can be turned around to give the result we want: $\mathbf{U}_{\mathrm{P}} = \mathbf{M}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Lambda}_{\mathrm{P}}^{-1} = \mathbf{g}^{-1}\mathbf{O}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{V}\boldsymbol{\Lambda}_{\mathrm{P}}^{-1}$. In deriving this result it has only been necessary to diagonalize a $P \times P$ matrix. Since P is of the order of tens for our two data sets, this is very easily accomplished and the covariance matrix is calculated directly from (3.26).

One more result is required before I show how all this formalism works out in practice. We can now calculate error-bars on a particular bin (and the covariances of each pair of bins as well, if required). But it would be very useful to be able to calculate things like the best estimate for $F(q)$ at any particular q and its expected error, or how much charge there is in a particular part on the nucleus, also with error estimates, or the nuclear mean square radius. All these quantities are linear functionals of the image.[15] Let the quantity we want to infer be Q. Then, by hypothesis, $Q = \sum_i a_i f_i$, for some coefficients a_i. For example, if we wanted to solve the last problem above, a_i would be taken to be equal to r_i^2. The mean value of Q over the posterior distribution is easily calculated:

$$\langle Q \rangle = \sum_i a_i \langle f_i \rangle, \tag{3.27}$$

and the squared error is almost as straightforward:

$$\begin{aligned}(\Delta Q)^2 &= \left\langle (Q - \langle Q \rangle)^2 \right\rangle \\ &= \sum_{ij} a_i a_j \,\mathrm{cov}(f_i, f_j)\end{aligned} \tag{3.28}$$

(Martin 1971).

If we are inferring the amount of charge in a **macrobin** I made up of several adjacent bins, as we shall be doing in the next section, the a_i for each bin in the macrobin are taken to be unity. Then $\langle Q \rangle = \sum_{i \in I} \langle f_i \rangle$ and $(\Delta Q)^2 = \sum_{ij \in I} \mathrm{cov}(f_i, f_j)$ which is a sum over a diagonal block of the covariance matrix.

A discussion of the method by which the results displayed in the next two sections were computed can be found in Appendix A.

[15] Indeed it is only linear functionals that represent observable physical properties at all (Skilling 1991).

3.6.2 $\mathbf{He^4}$

In Figure 3.6 are displayed estimates of the He^4 charge density at three different scales. Plotted for each scale are: the density which was displayed in Figure 3.2, which was made with 256 bins (smooth curve); the average density in each macrobin at the given scale (solid circles) and the estimated error in this average density (rectangles). The width of each rectangle shows the extent of the bin and the height is twice the expected error, centred on the most probable value. The bins have been amalgamated, from top to bottom, 4, 16 and 64 to one macrobin. The amount of charge in a macrobin is obtained by multiplying its most probable value by the width of the bin.

The increase in the error-bars with scale is very obvious. In the top figure, the errors are enormous—we are trying to infer very fine structure and so are being told that, with the given data, it is not possible to do that very well. On the coarser scale of the middle example, the errors have decreased in a plausible way, a trend which continues in the last plot. At the limit of one macrobin covering the entire nucleus, the error decreases to its minimum value, approximately that of the normalization point which was included at $q = 0$, since that single bin represents the average charge density of the nucleus and hence its total charge.

The 'shelf' feature in the reconstructed density between 3.75 and 4.75 fm is covered by the last bin in the last plot, the error of which is quite large enough to make this feature probably not significant. Similarly the last bin of the middle plot covers the region where the density just flirts with a change of sign, but it is completely swamped by the error on this bin.

Using the expressions (3.27) and (3.28) with $a_i = j(qr_i)$ for a range of qs, it is possible both to interpolate and to extrapolate the form factor. Indeed, because this quantity is an integral over the charge density, it is possible to infer the form factor anywhere and to calculate error-bands around it.[16] In Figure 3.7 is shown the inference of the form factor between 0 and $8\,\mathrm{fm}^{-1}$. The most probable form factor is shown as a solid line, the error-bands appear dashed. A number of features deserve comment. Note that the interpolation is everywhere sensible. Data points such as those at 3.9 and $4.5\,\mathrm{fm}^{-1}$ are very plausibly identified as being in error, lying as they do outside the 1σ error-bands. Where there are many data points with small errors falling on a smooth curve, such as below $3\,\mathrm{fm}^{-1}$, the form factor is determined very precisely: indeed the errors are in general smaller than those of the measured data because of the cumulative effect of all of the data containing mutually consistent information. In the region above $5.5\,\mathrm{fm}^{-1}$, where the data are not so constraining, observe how well the behaviour chimes with intuition. The reconstruction follows the data points closely and even the errors are reproduced. However, between two adjacent data

[16] The divergence of error-bands applies only to the density function f. The expected error of any integral of the charge density is, quite rightly, finite.

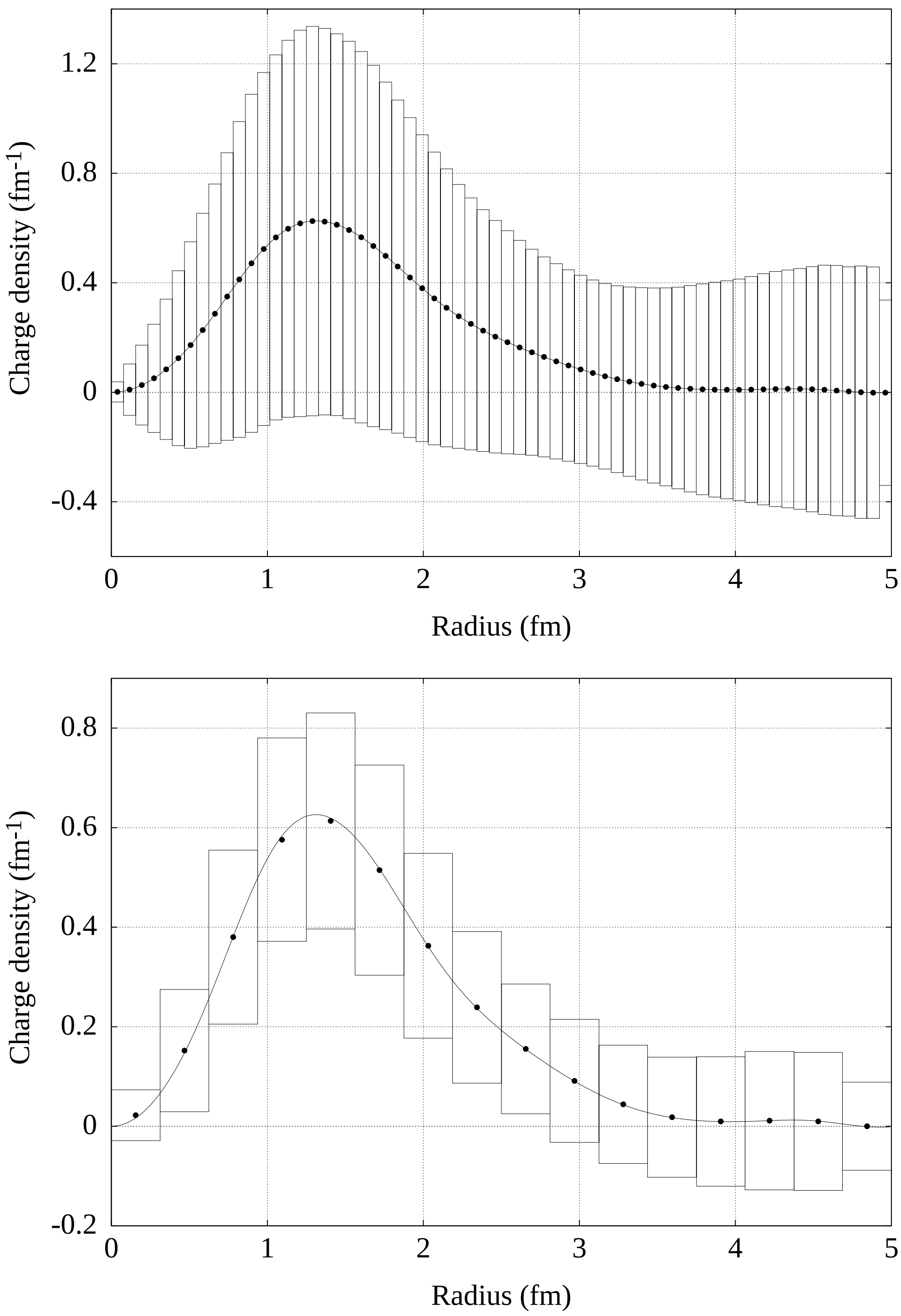


Fig. 3.6. Inferences about the He^4 charge density at three different scales. In each figure the smooth curve is the most probable density calculated with 256 bins, the dots represent the most probable density in the macrobins and the rectangles are the 1σ errors in those macrobins. The numbers of macrobins are 64 (top), 16 (bottom) and 4 (below).

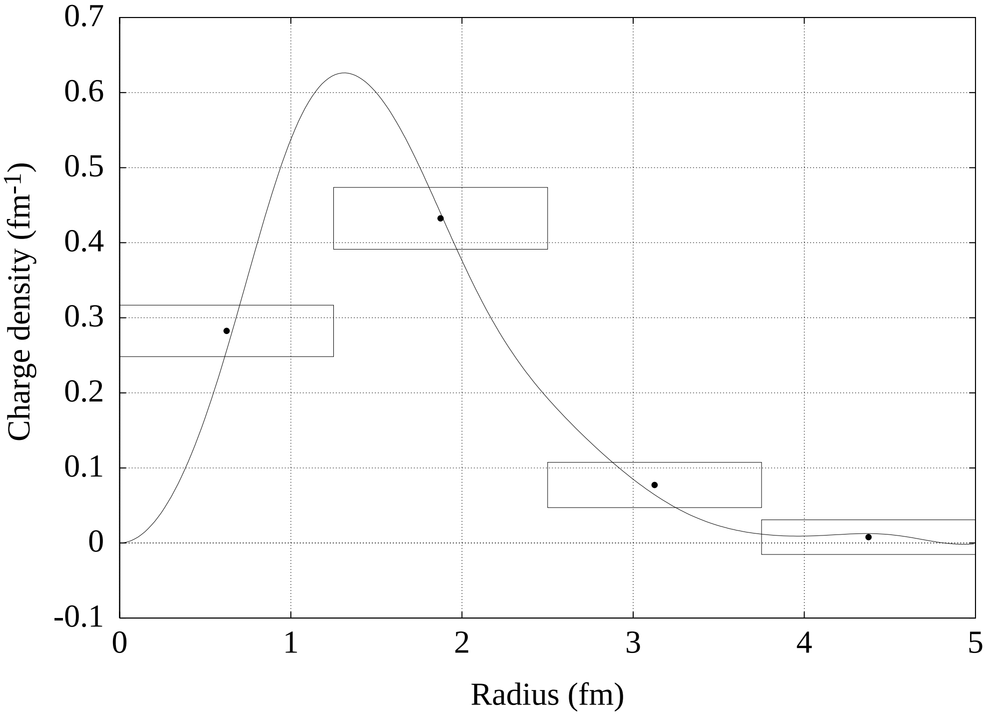


Fig. 3.6 cont.

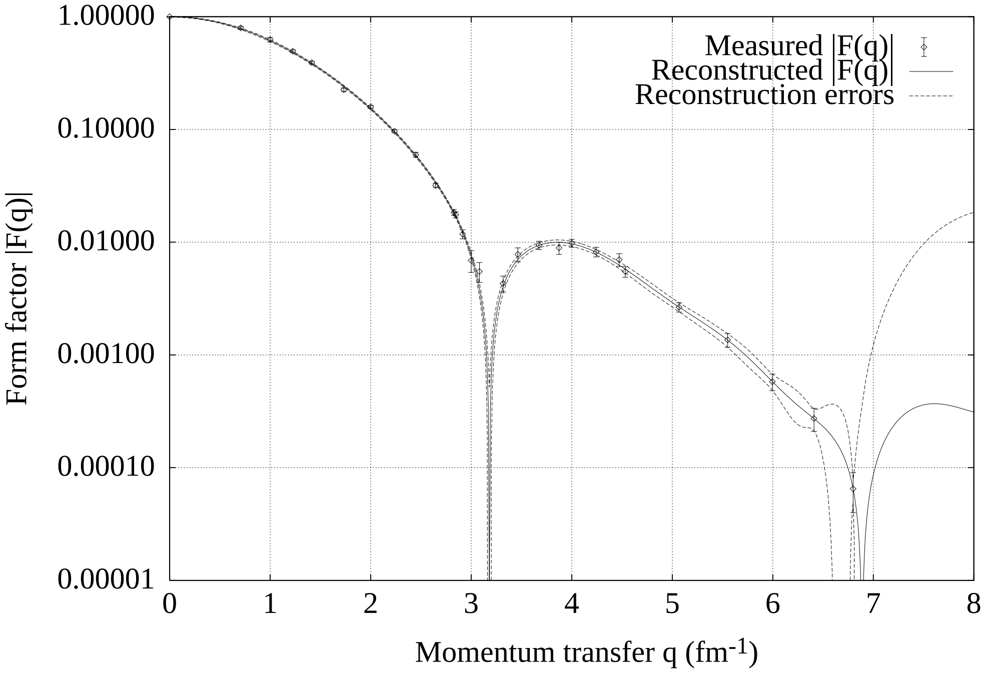


Fig. 3.7. The inferred He^4 form factor with error-bands. The most probable $|F|$ is shown (solid line) with 1σ error-bands (dashed lines). For comparison, the data is also plotted with its errors.

points (the points 6.0 and 6.4 fm^{-1}, for example), the interpolant becomes more uncertain, with error-bands increasing to a maximum roughly halfway between them.

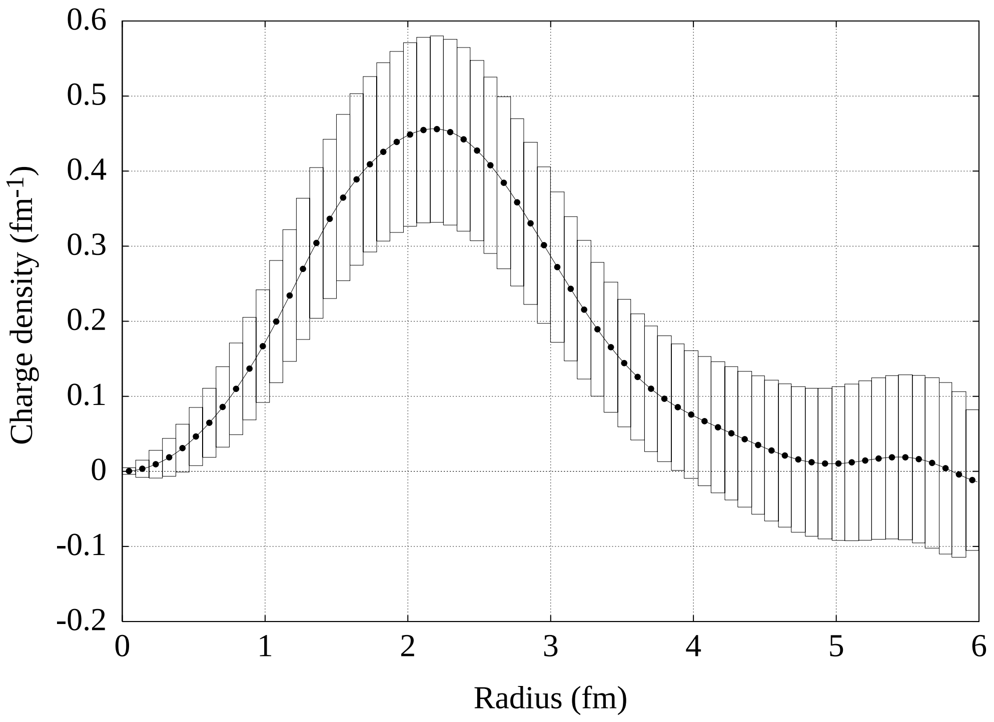
The extrapolant outside the range of the measured data gives a second diffraction minimum near 6.9 fm^{-1}. However the error-bands become very large very quickly making this a not terribly reliable prediction. We await new data with interest. It is not too surprising that the extrapolant becomes very uncertain so rapidly. The amount of physical prior information on the nucleus that went into our model is extremely small. If more information, that was *reliable*, could be added to the model, such as the range of nuclear forces, we would of course expect to obtain more predictive power. But for this problem at least, extrapolation was not the principle aim. Merely we sought to construct a good image of the charge density with a quantitative measure of its reliability. The remarkable thing about what has been done is that so much can be got for so little input, just the straight application of the rules of probability theory and some non-controversial assumptions.

3.6.3 C^{12}

For our other nucleus, the error-bands at different scales are shown in Figure 3.8 and the inferred form factor in Figure 3.9. The wiggles in the density above 4.5 fm are roughly on the scale of the middle diagram, the error-bars of which are large enough to render these features implausible. However the bottom figure shows that there is probably a not completely negligible amount of charge in this region, but there is not enough data to be able to infer its distribution.

3.7 Conclusion

I have attempted here to demonstrate an application of the Bayesian approach to inverse problems on a real example from nuclear physics, developing the necessary tools as they were needed. First a prior for a distribution of positive and negative charge was derived using a 'monkey argument'. This was used with a Gaussian likelihood to infer the most probable charge densities of the two nuclei He^4 and C^{12}. In order to quantify the reliability of these reconstructions, it was necessary to understand how to calculate error-bars. This revealed that how well-determined the charge density was depended on how closely one looked at it, and forced us to consider error-bars in a new light. I then showed that it was possible to calculate these quantities by doing computations only in a space the size of the number of data, which avoided the need directly to invert matrices of the size of the image. It turned out to be easy to calculate estimates with errors of any integral linear in the charge density, such as the form factor. All the



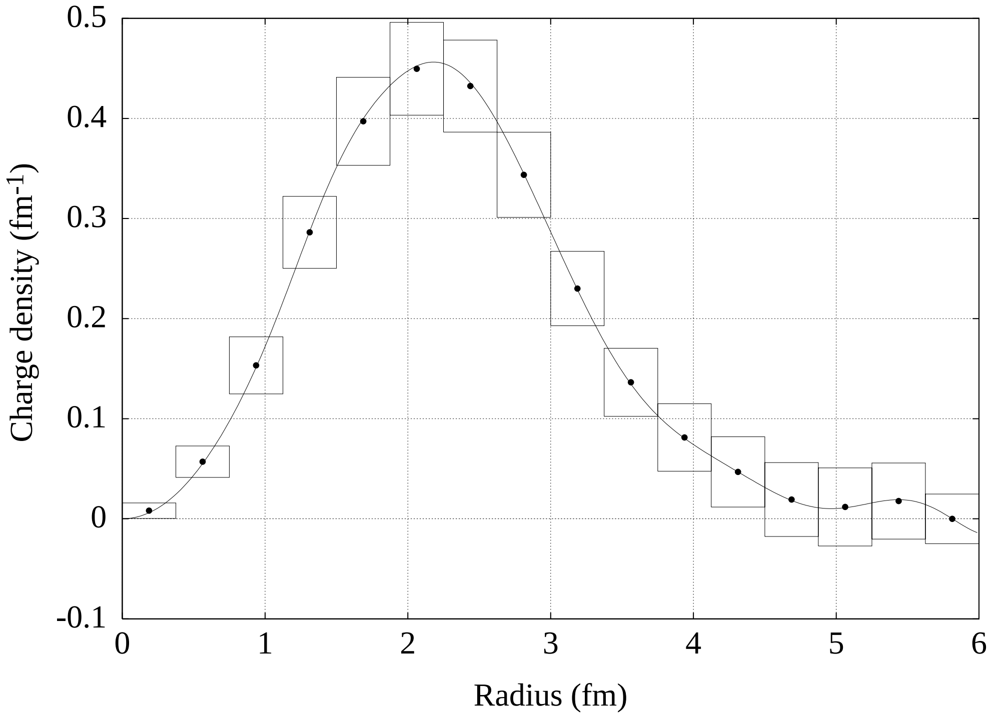


Fig. 3.8. Inferences about the C^{12} charge density at three different resolutions. See the matching diagram for He^4 for explanation.

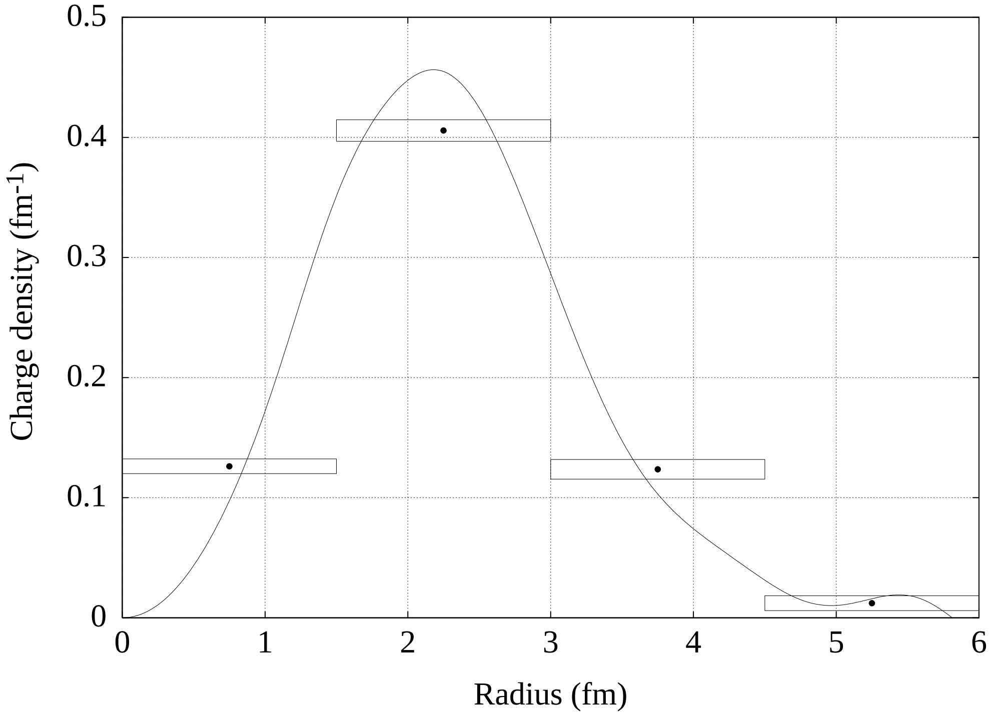


Fig. 3.8 cont.

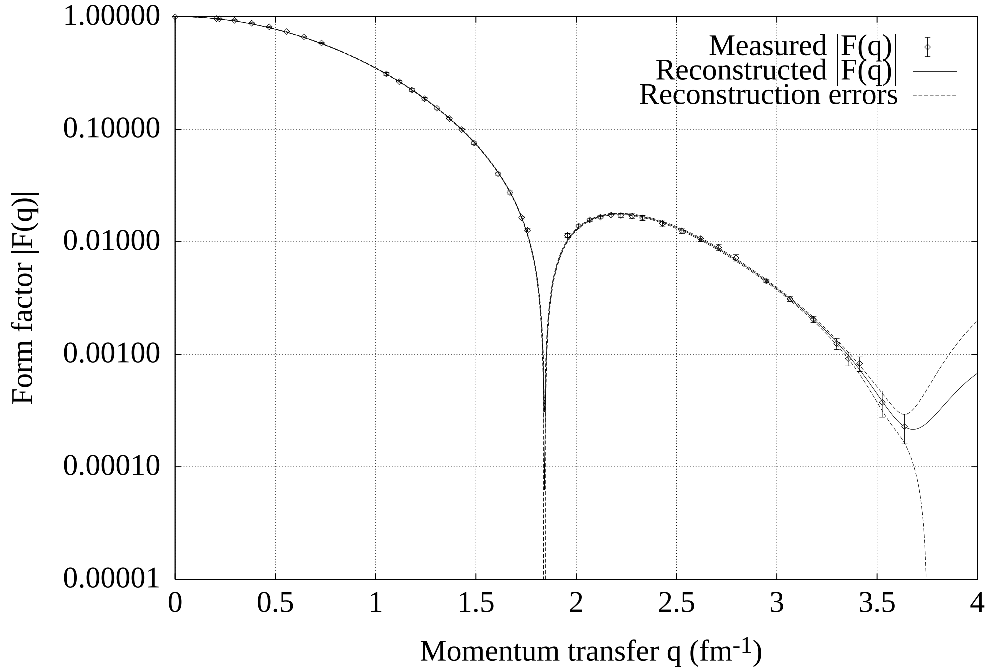


Fig. 3.9. The inferred C^{12} form factor with error-bands. See the matching diagram for He^4 for explanation.

inferences made chimed well with common sense, showing that elaborate models of the nuclear shape are not needed to recover the charge density.

References

Arnold, R. G., Chertok, B. T., Rock, S., Schütz, W. P., Szalata, Z. M., Day, D. *et al.* (1978). Elastic electron scattering from ^{3}He and ^{4}He at high momentum transfer. *Physical Review Letters*, **40**, 1429–32.

Buck, B. and Macaulay, V. A. (1992). Entropy and sunspots: their bearing on time-series. In *Maximum entropy and Bayesian methods, Seattle University, 1991* (ed. G. Erickson, P. Neudorfer and C. R. Smith). Kluwer, Dordrecht (in press).

Daniell, G. J. and Potton, J. A. (1989). Liquid structure factor determination by neutron scattering—some dangers of maximum entropy. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 151–62. Kluwer, Dordrecht.

Dreher, B., Friedrich, J., Merle, K., Rothhass, H. and Lührs, G. (1974). The determination of the nuclear ground state and transition charge density from measured electron scattering data. *Nuclear Physics*, **A235**, 219–48.

Frosch, R. F., McCarthy, J. S., Rand, R. E. and Yearian, M. R. (1967). Structure of the He^4 nucleus from elastic electron scattering. *Physical Review*, **160**, 874–9.

Gull, S. J. (1989). Developments in maximum entropy data analysis. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 53–71. Kluwer, Dordrecht.

Gull, S. J. and Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data. *Nature*, **272**, 686–90.

Jansen, J. A., Peerdeman, R. Th. and De Vries, C. (1972). Nuclear charge radii of ^{12}C and ^{9}Be. *Nuclear Physics*, **A188**, 337–52.

Lebedev, N. N. (1972). *Special functions and their applications.* Dover, New York.

Lindley, D. (1992). You've read the paper, now see the video. *Nature*, **358**, 97.

Macaulay, V. A. and Buck, B. (1990). Linear inversion by the maximum entropy method with specific non-trivial prior information. In *Maximum entropy and Bayesian methods, Dartmouth, U.S.A., 1989* (ed. P. Fougère), pp. 273–80. Kluwer, Dordrecht.

MacKay, D. J. C. (1992). Bayesian methods for adaptive models. Ph. D. thesis. California Institute of Technology.

Martin, B. R. (1971). *Statistics for physicists.* Academic Press, London.

Schiff, L. I. (1949). *Quantum mechanics.* McGraw-Hill, New York. Later editions: 1955, 1968.

Sick, I. and McCarthy, J. S. (1970). Elastic electron scattering from ^{12}C and ^{16}O. *Nuclear Physics*, **A150**, 631–54.

Skilling, J. (1989). Classic maximum entropy. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 45–52. Kluwer, Dordrecht.

Skilling, J. (1990). Quantified maximum entropy. In *Maximum entropy and Bayesian methods, Dartmouth, U.S.A., 1989* (ed. P. F. Fougère), pp. 341–50. Kluwer, Dordrecht.

Skilling, J. (1991). Fundamentals of MaxEnt in data analysis. In *Maximum entropy in action: a collection of expository essays* (ed. B. Buck and V. A. Macaulay), pp. 19–40. Clarendon Press, Oxford.

Skilling, J., Robinson, D. R. T. and Gull, S. F. (1991). Probabilistic displays. In *Maximum entropy and Bayesian methods, Laramie, Wyoming, 1990* (ed. W. T. Grandy, Jr and L. H. Schick), pp. 365–8. Kluwer, Dordrecht.

Watson, G. N. (1944). *A treatise on the theory of Bessel functions,* 2nd edn. Cambridge University Press. First edition: 1922.

4
Spectral estimation

I am thy father's spirit,
Doomed for a certain time
to walk the night.
Shakespeare: Hamlet

We discuss the Bayesian machine-tools which have been developed to solve different models of the spectral estimation problem, and add a powerful new one in the shape of a free-form model of the spectrum which is regularized entropically, and which can be used to reconstruct spectral density functions from noisy measurements of, for example, a time-series. It brings the usual benefits of noise suppression and artifact reduction, and has very wide applicability.

4.1 The problem

In a nutshell, spectral estimation is the fitting of harmonically varying functions (sines and cosines) to a data set in order to determine what frequencies are present, and possibly their amplitudes and phases.[1] This is a well-known inverse problem which has come of age in the era of the digital computer and for which a plethora of 'solutions' is available. There are different ways of formulating this apparently simple model of the data, that is, different questions we can ask, and each has its own Bayesian solution: we must be very clear what we want. For example, if we know how many periodicities there are in the data, we might want to estimate their parameters; alternatively, we might know that there is a small number of periodicities, and want to estimate that number; we might have reproducible measurements of the autocorrelation function of the signal at a few lags and want to estimate the power spectrum; or we might want to estimate the Fourier spectrum given noisy measurements of the signal. I shall briefly discuss the existing Bayesian solutions of two problems, spectral parameter estimation combined with model selection, and Burg's method, before going on to describe the noisy free-form problem and my solution of it.

[1] I assume here that the signal is measured as a function of time, i.e., that it is a time-series. However, most of what I shall say applies to spatial distributions just as well.

4.2 The parameter estimation approach

Bretthorst (1988) has worked extensively on the parameter estimation approach to the analysis of time-series. In his scheme, a typical model of the signal $f(t)$ might consist of N sinusoids each with its own frequency ν_n, amplitude R_n and phase ϕ_n, with $R_n \geq 0$ and $0 \leq \phi_n < 2\pi$:

$$\begin{aligned} f(t) &= \sum_{n=1}^{N} \big(R_n \cos(2\pi\nu_n t + \phi_n)\big) && (4.1) \\ &= \sum_{n=1}^{N} \big(A_n \cos(2\pi\nu_n t) + B_n \sin(2\pi\nu_n t)\big). && (4.2) \end{aligned}$$

The second form, with amplitude and phase replaced by the amplitudes A_n and B_n of two independent components of the signal is often more convenient: for the signal is a linear function of these amplitudes. Suppose that measurements of $f(t)$ are made at a discrete set of times $\{t_p : p = 1 \rightarrow P\}$ at each of which the datum might be modelled as containing an additive noise component. Symbolically, $D_p = f(t_p) + n_p$, where D_p is the pth datum and n_p is the noise in it, which is often assigned a Gaussian probability distribution. Once a prior on the parameters has been assigned, we can use the rules of probability theory, including Bayes' theorem and marginalization, to infer lots of things about the model. Since the data usually determine the parameters of small models quite well, much more accurately than any prior knowledge does, the prior is much less crucial than it is for image reconstruction problems. Typically, it will not vary much over the region of parameter space that is consistent with the data. Any broad, but normalizable, function will serve, presuming, of course, that we do not have very precise prior estimates of the parameters. The posterior distribution of the parameters is given by Bayes' theorem:

$$\begin{aligned} &\mathrm{prob}(\{A_n, B_n, \nu_n\}|\{D_p\}, N) = \\ &\qquad \frac{\mathrm{prob}(\{A_n, B_n, \nu_n\}|N)\,\mathrm{prob}(\{D_p\}|\{A_n, B_n, \nu_n\}, N)}{\mathrm{prob}(\{D_p\}|N)}; \qquad (4.3) \end{aligned}$$

all the usual games can be played with it. To estimate the frequencies, it is possible to marginalize away the amplitudes A and B analytically, given simple choices of the prior (for example, Gaussian in the amplitudes). For example, Bretthorst has shown that the probability density of the frequency in a single frequency model, given flat priors on the parameters, is proportional to the exponential of the Schuster periodogram divided by the noise variance.[2] Other objects which can be determined exactly are the conditional expectations of the amplitudes given the frequencies.

The model comparison stage of this technique is used to answer questions about the number N of sinusoids in the model. Consider a hypothesis space $\mathcal{N}$ in which there are models with different numbers of terms $\{N_\alpha\}$. Again, Bayes' theorem allows us to determine the posterior probability, but this time, of each model:

$$\mathrm{prob}(N_\alpha|\{D_p\},\mathcal{N}) = \frac{\mathrm{prob}(N_\alpha|\mathcal{N})\,\mathrm{prob}(\{D_p\}|N_\alpha)}{\mathrm{prob}(\{D_p\}|\mathcal{N})}. \tag{4.5}$$

If, *a priori*, models with different Ns are considered equally probable, then the ratio of the posterior probabilities of any two different models is

$$\frac{\mathrm{prob}(N_\alpha|\{D_p\},\mathcal{N})}{\mathrm{prob}(N_\beta|\{D_p\},\mathcal{N})} = \frac{\mathrm{prob}(\{D_p\}|N_\alpha)}{\mathrm{prob}(\{D_p\}|N_\beta)}. \tag{4.6}$$

Note that this is the ratio of the evidences (the denominator in (4.3)) for the two different models, each of which is evaluated from

$$\begin{aligned}\mathrm{prob}(\{D_p\}|N) \;=\; & \int \mathrm{prob}(\{A_n, B_n, \nu_n\}, \{D_p\}|N) \\ & \times\, \mathrm{d}^N\{A_n\}\,\mathrm{d}^N\{B_n\}\,\mathrm{d}^N\{\nu_n\}.\end{aligned} \tag{4.7}$$

These calculations, while straightforward in principle, are laborious in practice. For each model in turn, the most probable set of parameters must be identified, usually by a search in the parameter space. This is not a trivial task since the probability can be very sharply peaked indeed and contain many local maxima. After this has been done, the posterior must be approximated around the most probable point in order that the evidence may be calculated and hence different models compared. The trouble is that in this scheme there is no way to deal with models with lots of different Ns at the same time; each must be computed separately. The remedy is to change the model of the signal in order to capture its essential spectral features in one snapshot. We develop the necessary ideas in the following sections.

---

[2] The Schuster periodogram is defined by

$$C(\nu) = \frac{1}{P}\left|\sum_{p=1}^{P} D_p \exp(2\pi\mathrm{i}\nu t_p)\right|^2 \tag{4.4}$$

and is traditionally used to estimate the frequencies occurring in a time-series. For equally-spaced data, it is basically the squared modulus of the discrete Fourier transform evaluated at all frequencies.

4.3 Free-form spectral estimation

4.3.1 Formulating the free-form model

Instead of limiting ourselves to a model with a small number of frequencies, we might like to take the direction that was described in Chapter 2 and try to estimate a whole distribution of frequencies. This has been called the free-form approach (Skilling 1991). We can imagine taking (4.2) and generalizing it to an integral over spectral density functions $A(\nu)$ and $B(\nu)$:

$$f(t) = \int_0^{\nu_0} \big(A(\nu)\cos(2\pi\nu t) + B(\nu)\sin(2\pi\nu t)\big)\,\mathrm{d}\nu, \tag{4.8}$$

The frequency ν_0 is introduced because of ambiguities that arise if the intervals between data points are commensurate; this is discussed in §4.3.2 below. This representation of the signal is closely related to the Fourier transform

$$f(t) = \int_{-\nu_0}^{\nu_0} F(\nu)\exp(2\pi\mathrm{i}\nu t)\,\mathrm{d}\nu. \tag{4.9}$$

For, if $f(t)$ is real, then it is straightforward to show that $F(\nu) = F^*(-\nu)$ and hence that $A(\nu) = F(\nu) + F(-\nu)$ and $B(\nu) = F(-\nu) - F(\nu)$. As $\nu_0 \to \infty$, the usual expression for the Fourier transform is recovered. We know that the Fourier representations of a huge range of functions exist, and hence we can use (4.8) to describe most signals. Indeed, if we allow A or B to be generalized functions, nearly all "physicist's functions" can be so described.

Instead of having to carry around two separate functions A and B, it proves more convenient to replace them by a single function defined on twice the frequency range. For this purpose, the Hartley representation is ideal (Bracewell 1986). By defining the *cosine-and-sine* function $\mathrm{cas}(x) \equiv \cos(x) + \sin(x) = \sqrt{2}\cos(x - \frac{1}{4}\pi)$, we can write

$$f(t) = T\int_{-\nu_0}^{\nu_0} H(\nu)\,\mathrm{cas}(2\pi\nu t)\,\mathrm{d}\nu, \tag{4.10}$$

where the function $H(\nu)$ is called the **Hartley spectrum**. The amplitude information about a frequency $\tilde{\nu}$, say, previously stored in $A(\tilde{\nu})$ and $B(\tilde{\nu})$, is now contained in $H(\tilde{\nu})$ and $H(-\tilde{\nu})$. We can rewrite (4.10) in terms of amplitude and phase functions, $f(t) = T\int_0^{\nu_0} R(\nu)\,\mathrm{cas}\big(2\pi\nu t - \phi(\nu)\big)\,\mathrm{d}\nu$, then $R(\nu)$, the (Hartley) **amplitude spectrum**, equals $\sqrt{H^2(\nu) + H^2(-\nu)}$, and the phase function $\phi(\nu)$ is determined from $\cos\big(\phi(\nu)\big) = H(\nu)/R(\nu)$ and $\sin\big(\phi(\nu)\big) = -H(-\nu)/R(\nu)$. The amplitude spectra plotted in this thesis are scaled up by $\sqrt{2}$ and thus refer to the amplitude of a sinusoid. The total time interval T spanned by the data has been introduced as a scaling factor. This has been done in order to keep the inferred spectral function $H(\nu)$ of the same magnitude for data sets of different lengths. Longer data records allow better determinations of frequencies, which is manifested, for a free-form model, by narrower spectral peaks. However since the power in a peak is roughly preserved, then its height must increase. The T factor scales away this height increase. Consider an idealized signal $f(t) = \mathrm{cas}(2\pi\tilde{\nu}t)$, measured noiselessly everywhere between $-\frac{T}{2}$ and $\frac{T}{2}$. The most naïve assumption we can make is that it is zero outside this range, a familiar orthodox ploy. Then the Hartley inverse transform

$$H(\nu) = \frac{1}{T}\int_{-\infty}^{\infty} f(t)\,\mathrm{cas}(2\pi\tilde{\nu}t)\,\mathrm{d}t, \tag{4.11}$$

gives the inferred spectrum $H(\nu) = \mathrm{sinc}\,\pi(\nu-\tilde{\nu})T$. (The function $\mathrm{sinc}(x)$ is defined by $\sin(x)/x$ $(x \neq 0)$; 1 $(x = 0)$.) The value of $H(\tilde{\nu})$, namely unity, is independent of T, although the width of the sinc function decreases inversely with T.

A result which will be required later is the transformation that the spectrum suffers when the time coordinate is shifted, by τ say. This is most simply written in terms of the amplitude and phase, $R(\nu)$ and $\phi(\nu)$. Let the datum labelled with the time t be rechristened with a new label $t' = t-\tau$; let the signal as a function of the new variable be $f'(t')$, which by definition equals $f(t)$. The problem is to find the functions $R'(\nu)$ and $\phi'(\nu)$ that appear in the Hartley representation of $f'(t')$, namely, $T\int R'(\nu)\,\mathrm{cas}\big(2\pi\nu t' - \phi'(\nu)\big)\,\mathrm{d}\nu$, in terms of $R(\nu)$ and $\phi(\nu)$. By equating the amplitudes and phases on each side of the equality $f'(t') = f(t)$ at a particular ν, bearing in mind the definition of t', it is easy to see that $R'(\nu) = R(\nu)$ and $\phi'(\nu) = \phi(\nu) - 2\pi\nu\tau \bmod 2\pi$.

In the next two sections, I describe what limitations arise from the finiteness of the data. These are determined by the two time scales inherent in the data, namely the total time spanned by the data and the time intervals between data points.

4.3.2 Aliasing

There is a fundamental ambiguity that arises in the spectral analysis of a signal when all the time-intervals between data points are commensurate. It turns out that there are an infinite number of sinusoids of different frequencies that can fit the signal *at the measured points* equally well (Figure 4.1). Unless there is prior information which enables one to pick out one of these frequencies as more probable, the inverse problem does not have a unique solution. Consider the component of the signal which has frequency $\tilde{\nu}$ in the Hartley representation, namely, $\mathrm{cas}(2\pi\tilde{\nu}t)$. If the intervals between the data points are commensurate, we can write the time of the pth point as $t_p = \tau n_p + \tau_0$, where τ is the longest time interval of which all others are a multiple, τ_0 is an offset and the n_ps are integers with no common factor.

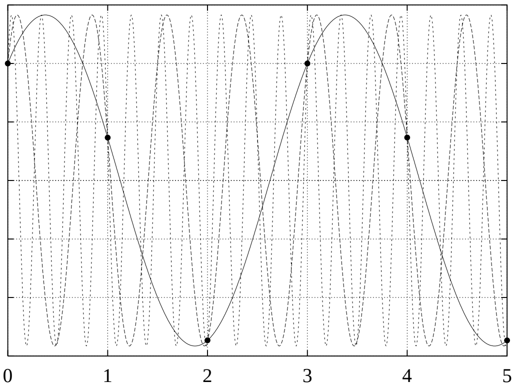


Fig. 4.1. The ambiguity inherent in inferring a spectrum from data points spaced commensurately is indicated. Each of three sinusoidal signals of frequencies 0.33, 1.33 and 3.33 fit the, in this case unit-spaced, data points (solid circles) in exactly the same way. It is impossible to choose between them in the light of the data alone.

Then, for any pair of data points, $t_p - t_q = \tau(n_p - n_q) = \text{integer} \times \tau$, as required. Any cas functions with frequencies ν different from $\tilde{\nu}$ which have the same values at *all* the times t_p as $\mathrm{cas}(2\pi\tilde{\nu}t)$ must satisfy the equation $\mathrm{cas}(2\pi\nu t_p) = \mathrm{cas}(2\pi\tilde{\nu}t_p)$. This implies that $(2\pi\nu t_p = 2\pi\tilde{\nu}t_p) \bmod 2\pi$, and hence that

$$\tau_0(\nu - \tilde{\nu}) = n_p\tau(\tilde{\nu} - \nu) \mod 1. \tag{4.12}$$

In the case that $\tau_0 = 0$, the LHS is zero. Therefore, since n_p is an integer, $\tau(\tilde{\nu} - \nu)$ must be an integer also and

$$\nu = \tilde{\nu} + \frac{m}{\tau} \qquad \text{for} \qquad m \in \mathbb{Z}. \tag{4.13}$$

Other cases can be made to fall into the $\tau_0 = 0$ category by shifting the time origin. Thus the frequencies which are entirely equivalent from the point of view of the data are uniformly separated on the real line with a spacing τ^{-1}; in the absence of specific prior knowledge, the spectrum $H(\nu)$ is a periodic function of ν.[3] The amplitude function, for this kind of data, has an additional symmetry. From above, the amplitude at ν, $\sqrt{H^2(\nu) + H^2(-\nu)}$ equals $\sqrt{H^2(\nu) + H^2(m\tau^{-1} - \nu)}$. But the amplitude at $\nu = m\tau^{-1} - \nu$ clearly has the same value, that is, the amplitude spectrum has reflectional symmetry about each $\nu = \frac{1}{2}m\tau^{-1}$, as well as the above translational symmetry.

The largest unambiguous range of frequencies has a length $\frac{1}{2}\tau^{-1}$. In the Hartley spectrum, both the phase and the amplitude at each frequency have to be stored and hence the range is twice as long. If a data set, sampled

[3] This is true for $\tau_0 = 0$. In other cases, the amplitude spectrum has this periodicity, but the phases are changed.

like this, has not been filtered to restrict power to one particular frequency window, then any power outside that region will be translated or reflected into the region. This is the phenomenon of **aliasing**. Even if filtering has not been done, either because the data have been taken often enough to capture structure up to frequencies so high that there is a good physical reason to suppose that no higher frequencies could be present, or merely because of an oversight, there is nothing to be gained from reconstructing the Hartley spectrum outside of an unambiguous frequency window. This might as well be taken as $-\nu_0 < \nu \leq \nu_0$ where $\nu_0 = \frac{1}{2}\tau^{-1}$. This is why I put the limits I did on the integrals defining the image-to-data transformation, for example (4.10). For data equally-spaced and time measured in units such that $\tau = 1$, the value of ν_0 is $\frac{1}{2}$. Thus, for example, data points measured yearly are ambiguous about periodicities shorter than half of one year.

4.3.3 Truncation and sinc-wiggles

In our discussion of inverse problems in Chapter 2, it was repeatedly pointed out that, when a Gaussian likelihood is used, the kernel and its eigenstructure play a vital rôle in determining the most probable reconstruction. We will now look at that again in the context of spectral estimation. Let us suppose that tha data points are equally spaced, each with an expected error σ and let the time scale be chosen so the points are symmetrically disposed about $t = 0$, and their separation τ is unity. Again we use ket notation, with, in the frequency representation, $H(\nu) = \langle \nu \mid H \rangle$ and $\mathrm{cas}(2\pi\nu t_p) = \langle \nu \mid p \rangle$, so that $f(t_p) = \langle p \mid H \rangle$.[4] In this representation, the kernel $\hat{K}$ equals $K(\nu, \nu') = \sigma^{-2} \sum_p \langle \nu' \mid p \rangle \langle p \mid \nu \rangle = \sigma^{-2} \sum_p \mathrm{cas}(2\pi\nu t_p) \times \mathrm{cas}(2\pi\nu' t_p)$, where t_p takes on the values $-\frac{1}{2}(P-1), -\frac{1}{2}(P-1)+1, \ldots, \frac{1}{2}(P-1)$. Performing the sum gives

$$K(\nu, \nu') = \frac{1}{\sigma^2} \frac{\sin \pi P(\nu - \nu')}{\sin \pi(\nu - \nu').} \tag{4.14}$$

The eigenfunctions are easy to determine. For the bracket $\langle p \mid p' \rangle$ equals $\delta_{pp'}$, and hence $\hat{K} \mid p \rangle = \sigma^{-2} \sum_{p'} \mid p' \rangle \langle p' \mid p \rangle = \sigma^{-2} \sum_{p'} \mid p' \rangle \delta_{pp'}$, which equals $\sigma^{-2} \mid p \rangle$, if t_p is one of the times labelling the data points, and zero, otherwise. Thus the normalized eigenkets $\mid g_p \rangle$ are the set of $\mid p \rangle$s and the eigenvalues are σ^{-2}, if $-\frac{1}{2}(P-1) \leq t_p \leq \frac{1}{2}(P-1)$ (the measured space) and zero, otherwise (the null space). The datagram is easily shown to be $\sum_p D_p \, \mathrm{cas}(2\pi\nu t_p)$.

Consider the somewhat idealized spectrum consisting of a delta function of unit strength at $\nu = \tilde{\nu}$; the signal corresponding to this is $f(t) = \int \delta(\nu - \tilde{\nu}) \, \mathrm{cas}(2\pi\nu t) \, \mathrm{d}\nu = \mathrm{cas}(2\pi\tilde{\nu} t)$. Suppose the signal $f(t)$ was measured

[4] The T factor in (4.10) will be ignored in this section.

at P times. The datagram for this data set is $\sum_p \operatorname{cas}(2\pi\tilde{\nu}t_p)\operatorname{cas}(2\pi\nu t_p)$ which equals $\sin\bigl(\pi P(\nu-\tilde{\nu})\bigr)/\sin\bigl(\pi(\nu-\tilde{\nu})\bigr)$, a function which also appeared above and is familiar from the theory of diffraction. For P odd, it is periodic in $\nu-\tilde{\nu}$ with a period of one (see §4.3.2) and has by far its largest maximum when $\nu=\tilde{\nu}$ (and at the period one repeats) when its value is P. On either side of this main peak, rapidly decreasing oscillations fall away. These are **sinc-wiggles**. The first zeros on either side of the main peak are at $|\nu-\tilde{\nu}|=(2P)^{-1}$, so that the width of the main spike is inversely proportional to the number of data points and hence the total time spanned by the data.

The broadening of the ideal spectrum—the delta function—into the diffraction function arises precisely because the data is truncated. Indeed sinc-wiggles are the signatures of truncation. The datagram displays this problem at its worst because it makes a very bad prediction of the signal outside the interval where it was measured, namely that it is zero, which is completely at odds with the true behaviour. This is a feature of the free-form model; it does not have built into it the fact that the signal is believed to behave in much the same way outside the time interval when it was measured as it does in that interval, which is the case with the discrete model of §4.2. It is the price to be paid for seeking a more flexible approach. A proper free-form theory, such the one developed in §4.5, will do somewhat better than the datagram at extrapolating the data and will be able largely to suppress sinc-wiggles, but the finite width of spectral lines will remain.

4.4 Burg's method

Somewhat parenthetically, I shall briefly discuss a rather specialized free-form method that is due to Burg (1975), the Bayesian justification of which has been elucidated by Jaynes (1982, 1985). This method assumes that data consisting of reproducible and essentially *noiseless* measurements of the autocorrelation of a fluctuating time-series at a number of lags is available. These are used as testable information to assign the probability of the signal from which they derived by an application of the principle of maximum entropy (see Chapter 1). From this distribution it is not hard to derive an expression for the expected power spectrum

$$P(\nu)=|F(\nu)|^2=\Bigl(\sum_k \lambda_k \exp(2\pi i\nu k)\Bigr)^{-1}, \tag{4.15}$$

where the λs are Lagrange multipliers that are introduced to incorporate the autocorrelation constraints, assumed measured at integer lags.

This expression can have poles and hence is very well suited to reconstructing spiky spectra. It is able to extrapolate the autocorrelation function beyond where is was measured, in contrast to the widely used Blackman–Tukey method (Blackman and Tukey 1958), which predicts all unmeasured autocorrelations to be zero, and, what is more, weights the measured autocorrelations so that they go smoothly to zero for the largest lags! Although Burg's method is a pretty theory which has some very appealing properties, it provides the solution to a rather rarely occurring problem: noiseless autocorrelations do not crop up every day. It is much more common to have available a time-series itself, contaminated with an additive noise component. It was this kind of situation that was described in §4.2 but with a very specific model of the signal.

In the following section, I shall develop the solution of a spectral estimation problem, where the measured signal is noisy, like Bretthorst's, but the inference is a spectral density function, somewhat like Burg's.

4.5 Invariance and a prior

We have reduced the problem of spectral estimation to the kind of inverse problem we looked at in Chapter 2, with (4.10) playing the rôle of the image-to-data transformation. So the game is to infer $H(\nu)$ by assigning a prior and a likelihood, which are used in Bayes' theorem. The likelihood is no different, in principle, from usual and we shall happily assign it to be Gaussian.

In trying to formulate a prior for the spectrum $H(\nu)$, it is necessary to think rather carefully about what we know of this function. It is not natural to think of $H(\nu)$ and $H(-\nu)$, for each ν, as independent *a priori*, for they both are giving information about the component of the signal at that frequency. In other words, we would, in general, be rather surprised if structure appeared in $H(\nu)$, for some ν, but there was none at $-\nu$. That is not to say that this could not happen: for components of certain phases, it surely could. However, such situations are the exception rather than the rule. In general, finding that there was structure at frequency $\tilde{\nu}$ should change our probability for structure at $-\tilde{\nu}$. In symbols,

$$\mathrm{prob}(H^+|H^-) \neq \mathrm{prob}(H^+), \tag{4.16}$$

where, for any $\tilde{\nu}$, $H^+ = H(\tilde{\nu})$ and $H^- = H(-\tilde{\nu})$. However, although the probability $\mathrm{prob}(H^+, H^-)$ should be correlated, knowledge of H^+ does not, in general, give us any idea of the *sign* of H^-, or *vice versa*. Thus we might expect that $\mathrm{prob}(H^+, H^-)$ is a function only of ${H^+}^2$ and ${H^-}^2$. This will be the case.[5]

[5] Notice that the unconditional prior expectations of H^+ and H^- are zero, as is their covariance and hence their correlation coefficient (see, for example, Martin 1971). This reveals the interesting fact that a vanishing correlation coefficient does not imply that the two variables involved have independent probabilities, although it is true that a pair of independent variables have a vanishing correlation coefficient.

We can conclude that any prior which is independent in all the bins, as was the case in Chapter 3 for inferring charge densities, is ruled out (Buck and Macaulay 1992). However there is nothing to stop us assigning a prior which assumes no correlation between the spectral components of different periodicities. It is just that the pair of numbers that characterize the spectrum at each frequency are not naturally assigned independent distributions. Suppose we discretize the frequency range $0 \le \nu \le \nu_0$ into N bins labelled by n, then the probability we are trying to construct is $\mathrm{prob}(\{H_n^+, H_n^-\})$, where $H_n^\pm = H(\pm\nu_n)$. By the assumption of no inter-frequency correlation this simplifies to $\prod_{n=1}^N \mathrm{prob}(H_n^+, H_n^-)$ or equivalently, $\prod_{n=1}^N \mathrm{prob}(R_n, \phi_n)$ in the amplitude/phase form, which proves more convenient for this discussion. Let the probability density of R_n and ϕ_n be $g_n(R_n, \phi_n)$. Then

$$\mathrm{prob}\big(\{[R_n, R_n+\mathrm{d}R_n], [\phi_n, \phi_n+\mathrm{d}\phi_n]\}\big) = \prod_{n=1}^{N} g_n(R_n, \phi_n)\,\mathrm{d}R_n\,\mathrm{d}\phi_n. \quad (4.17)$$

We can get further by recognizing that there is a useful invariance which is implicit in the nature of the problem. Jaynes has shown repeatedly (Jaynes 1968, 1973, in preparation) how transformations which map the problem into another one about which our state of knowledge is the same can be used to constrain the form of the prior distribution. The method relies on Jaynes' consistency principle, which can be stated as

> "In two problems where we have the same prior information, we should assign the same prior probabilities."

In our problem, the essential thing to recognize is that the time that is attached to any data point is only a label. In particular, the choice of the time origin is entirely arbitrary. Thus whether I say that time 0 was `16:18.03 BST` yesterday or 1992 years ago is irrelevant to the inferences we should make. Recognizing that this is so, we must insist that the prior is insensitive to this choice. Therefore, applying Jaynes' principle, we equate the prior probabilities of the spectra corresponding to two different time origins:

$$\prod_n g_n(R_n, \phi_n)\,\mathrm{d}R_n\,\mathrm{d}\phi_n = \prod_n g_n(R'_n, \phi'_n)\,\mathrm{d}R'_n\,\mathrm{d}\phi'_n, \quad (4.18)$$

for all τ. As in §4.3.1, a prime indicates that a quantity has a value appropriate to the time origin having been shifted forward by τ. But the relationships between the primed and the unprimed expressions were given earlier:

$$\begin{aligned} R'_n &= R_n, & (4.19)\\ \phi'_n &= \phi_n - 2\pi\nu_n\tau \mod 2\pi. & (4.20) \end{aligned}$$

It is easy to calculate the probabilities of the unprimed variables that are implied by the RHS of (4.18), using the usual technique (e.g., Martin 1971). Since the expression relating each pair (R_n, ϕ_n) to (R'_n, ϕ'_n) is one-to-one, with a Jacobian determinant that is unity, we have

$$\prod_n g_n(R'_n, \phi'_n)\,\mathrm{d}R'_n\,\mathrm{d}\phi'_n = \prod_n g_n(R_n, \phi_n - 2\pi\nu_n\tau)\,\mathrm{d}R_n\,\mathrm{d}\phi_n. \qquad (4.21)$$

But now we can equate the LHS of (4.18) to the RHS of (4.21) and, since this relation must hold for each (R_n, ϕ_n) pair, the function g_n must satisfy

$$g_n(R_n, \phi_n) = g_n(R_n, \phi_n - 2\pi\nu_n\tau) \qquad (4.22)$$

for *any* τ. But this freedom in τ means that the second argument on the RHS of (4.22) can take on any value in the range $[0, 2\pi)$, while the RHS is unchanged. Hence g_n cannot be a function of that second argument:

$$g_n(R_n, \phi_n) = h_n(R_n). \qquad (4.23)$$

This is the central result of this section. It necessitates the prior being only a function of the amplitude at each frequency. We can argue further that there is nothing special about any of the frequencies ν_n and so there is no reason to assign any of the functions h_n to be different from any other. So we drop the n on h and write

$$\mathrm{prob}\big(\{[R_n, R_n + \mathrm{d}R_n], [\phi_n, \phi_n + \mathrm{d}\phi_n]\}\big) = \prod_{n=1}^{N} h(R_n)\mathrm{d}R_n\,\mathrm{d}\phi_n, \qquad (4.24)$$

which is to be compared with our starting equation (4.17).

The choice of the function $h(x)$ is still left open by this argument. We saw in earlier chapters, however, that an entropic form has many advantages, and so the selection I shall make will be from that family. The usual $-x \ln x$ form has the disadvantage that its curvature is not negative everywhere in the parameter space, so that a unique maximum in the posterior cannot be guaranteed. In fact such a prior has a 'dimple' at $R_n = 0$. There is nothing wrong in principle with that, but it could lead to computational difficulties in finding the global maximum. Instead I propose to use the 'sinh' form that was derived in Chapter 3, which does have a strictly negative curvature. That is

$$h(R_n) \propto \exp\Big(\alpha\Big[\sqrt{R_n^2 + m^2} - m - R_n \sinh^{-1}\frac{R_n}{m}\Big]\Big), \qquad (4.25)$$

when $R_n \geq 0$ and $h(R_n) = 0$ for $R_n < 0$.

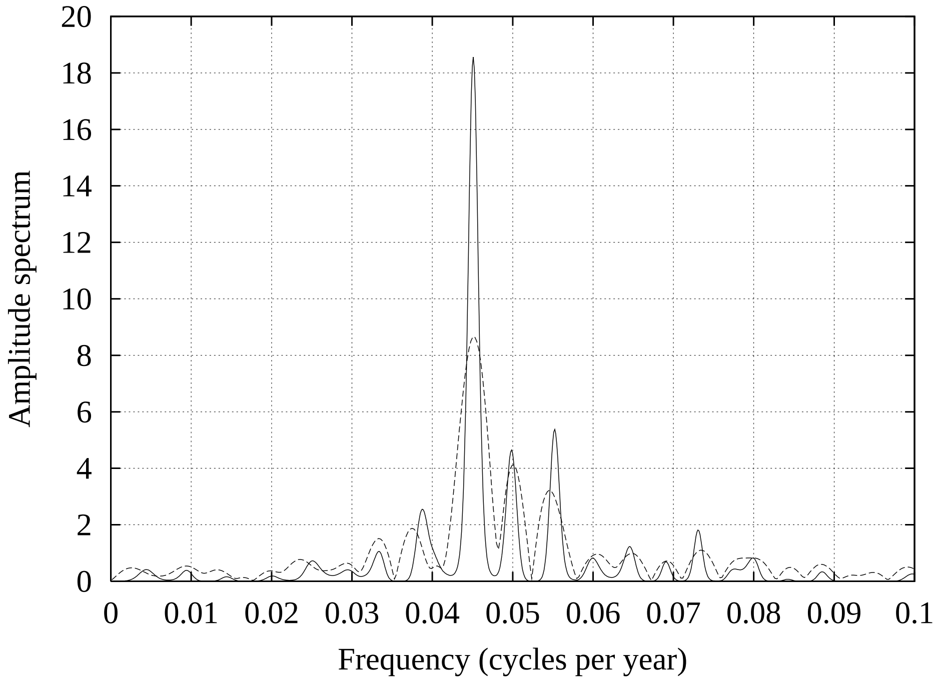


Fig. 4.2. The amplitude spectrum of the derectified sunspot data of Chapter 5 restored entropically (solid line) compared with an orthodox spectrum (dashed line) derived from the Schuster periodogram. Notice the sharper spectral peaks, the improved resolution and the cleaner baseline of the Bayesian reconstruction.

The most probable spectrum is found by maximizing the posterior probability, which is proportional to the prior multiplied by the likelihood. A computational device aids in this maximization: to avoid searching on the prior's natural variables $\{R_n, \phi_n\}$, which have awkward restrictions on their possible values ($R_n \geq 0$ and $\phi_n \in [0, 2\pi)$), it is much better to reexpress the posterior in terms of the variables $\{H_n^+, H_n^-\}$ and to search on these, since they are unrestricted.

In order to show the resolution enhancement and noise suppression of an entropically restored spectrum, I plot in Figure 4.2 the amplitude spectrum of the derectified sunspot data (see Figure 5.4), which we shall look at in a lot more detail in Chapter 5. Here I use it just as an example to show how much cleaner is the entropic reconstruction than an orthodox spectral estimator such as the amplitude of the Schuster periodogram, which is shown as a dashed curve. It should not be forgotten that the effect of the entropy is partly to *smooth* the spectrum. Unrealistically sharp features are suppressed because they have an entropy very much less than maximal. The fact that the orthodox spectrum produces *less* well defined features shows that it is implicitly assuming a very bad model of the signal. For example, it extrapolates unrealistically and it assumes that all of each data point is signal and no part is noise.

4.5.1 Recovering a discrete model

It is impossible with a free-form model and noisy data to recover the infinitely sharp delta functions in the spectrum that are the frequency domain representations of discrete periodicities in the signal. The entropy will always strongly oppose such features. However it is tempting to identify any sharp peaks in an inferred spectral density function as being signs of these periodicities that have been broadened by the finiteness of the data record and perhaps by noise. We can recover a model with a discrete number of lines by integrating over the peaks in the spectrum. Suppose that there is a spectral feature which we feel sure represents a discrete periodicity. The contribution that it makes to the signal is

$$f_{\mathrm{peak}}(t) = T \int_{\mathrm{peak}} H(\nu) \operatorname{cas}(2\pi\nu t) \, \mathrm{d}\nu. \tag{4.26}$$

However, since $H(\nu)$ is sharply-peaked inside the region of integration, around the frequency ν_k say, we can approximate $\operatorname{cas}(2\pi\nu t)$ as $\operatorname{cas}(2\pi\nu_k t)$ in the integrand, and write

$$f_{\mathrm{peak}}(t) \approx T \operatorname{cas}(2\pi\nu_k t) \int_{\mathrm{peak}} H(\nu) \, \mathrm{d}\nu. \tag{4.27}$$

This is a single cas function with an amplitude determined by an integration over the spectrum. In general, a feature at $-\nu_k$ will be present as well, which can be integrated to give the amplitude of $\operatorname{cas}(-2\pi\nu_k t)$, and hence the overall amplitude and phase of the component of the signal of frequency ν_k can be recovered.

4.6 Conclusion

In this chapter, different approaches to spectral estimation have been discussed. A way to infer the spectrum of a noisy time-series, by using Bayes' theorem with a prior entropic in the spectrum's amplitude and uniform in its phase, has been developed and illustrated. In the final chapter of this thesis, I shall apply this method to an investigation of the spectral features of the Wolf sunspot record.

References

Blackman, R. B. and Tukey, J. W. (1958). *The measurement of power spectra.* Dover, New York.

Bracewell, R. N. (1986). *The Hartley transform.* Oxford University Press, New York.

Bretthorst, G. L. (1988). *Bayesian spectral analysis and parameter estimation* (Lecture notes in statistics, Vol. 48). Springer-Verlag, New York.

Buck, B. and Macaulay, V. A. (1992). Entropy and sunspots: their bearing on time-series. In *Maximum entropy and Bayesian methods, Seattle University, 1991* (ed. G. Erickson, P. Neudorfer and C. R. Smith). Kluwer, Dordrecht (in press).

Burg, J. P. (1975). Maximum entropy spectral analysis. Ph. D. thesis. Stanford University.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, **SSC-4**, Sept., 227–41.

Jaynes, E. T. (1973). The well-posed problem. *Foundations of Physics*, **3**, 477–92.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, **70**, 939–52.

Jaynes, E. T. (1985). Where do we go from here? In *Maximum-entropy and Bayesian methods in inverse problems* (ed. C. R. Smith and W. T. Grandy, Jr.), pp. 21–58. Reidel, Dordrecht.

Jaynes, E. T. (in preparation). *Probability theory: the logic of science.* Fragments available from the author: Wayman Crow Professor of Physics, Washington University, St. Louis, MO 63130, USA.

Martin, B. R. (1971). *Statistics for physicists.* Academic Press, London.

Skilling, J. (1991). On parameter estimation and quantified MaxEnt. In *Maximum entropy and Bayesian methods, Laramie, Wyoming, 1990* (ed. W. T. Grandy, Jr. and L. H. Schick), pp. 267–73. Kluwer, Dordrecht.

5

A new look at the sunspot cycle

One saith the sun stands, another he moves;
a third ... finds certain spots or clouds in the sun.
Robert Burton: The Anatomy of Melancholy

In this chapter the formalism that was described in Chapter 4 is applied to an analysis of the sunspot index of Wolf over the last 300 years. The cleanness of the spectra provided by the method allows economical models of the solar activity cycle to be identified. These spectra confirm features of this data set which have been identified by others in the past, but they also show up the need for new twists, some of which are expounded.

5.1 Introduction

The first telescopic observations of the dark patches on the surface of the sun were made around 1610 by Fabricius and Galileo (Foukal 1990; Coyne 1991). Twenty years later it was recognized that their motion across the solar disk was revealing the rotation of the sun (at a period of $\sim$27 days) and the tilt of its axis to the ecliptic.[1] Between 1645 and 1715, there may have been a solar activity minimum, very few spots being visible, which has been much argued about and named after Maunder who, with Spörer, rediscovered it in solar records in the late 19th century (Eddy 1976).

It was Schwabe in 1843 who discovered the 11 year cycle in the sunspot record and Wolf who recognized a correlation between peaks in the cycle and geomagnetic storms. The latter's tabulation of his **Sunspot Index** began in 1849. This index, somewhat arbitrarily defined in terms of the number of spots f, the number of groups of spots g and an observer dependent scale-factor k, by $k(10g + f)$, has been shown to correlate well with other more objective measures of solar activity such as the total area covered by sunspots (Waldmeier 1961). The index back to 1700 has been inferred from contemporaneous descriptions of solar observations; current values are available in, for example, *Sky and Telescope*. The index fluctuates vigorously on time-scales of days and weeks due to the rotation of

[1] In passing it should be noted that the sun does not rotate as a rigid body; the angular velocity at the equator is larger than at the poles.

the sun and the short life of individual spots, but over months and years an overall cyclical behaviour is very obvious. It is the fine structure of the record on the scale of years with which we are concerned here. To investigate these longer term features it is convenient to average the data over these periods and to work with yearly or biyearly means. This has the spin-off of reducing the amount of data to be analysed.

An important feature in the sunspot record which has been recognized more recently, by Sonett (1983), is the **Great Solar Anomaly** (GSA). This occurred between approximately 1770 and 1820, for which period the index shows large variations from the norm both in amplitude and phase. (It is clearly visible in Figure 5.2 near 1800.) We shall be showing how the spectral features of the record change depending on the treatment given this anomaly.

There is one important reinterpretation of the data which it turns out to be vital to make. It is suggested by Hale's identification early this century of a 22 year magnetic solar cycle; this is related to the spot cycle as follows. Spectroscopic studies have indicated that a magnetic field of $\sim$0.1 T traverses a typical spot. Now, spots often appear in pairs near the same latitude with fields of opposite polarities. Strikingly, the polarities of the leading spots—those which appear 'in front' in the motion of the spots across the disk—are the same *within* each hemisphere, but *opposite* from the polarities characterizing the other hemisphere. The magnetic cycle is manifest in the alternation of the polarities of these leading spots between cycles of the sunspot numbers. Figure 5.1 shows one model—the Babcock model—of the structure of this magnetic field. This suggests that the sunspot cycle in fact represents a rectified realization of the magnetic driving cycle. Since the interesting physics will lie in understanding the underlying dynamics of the forcing mechanism, it would seem useful to *derectify* the measured index in order to get back to the magnetic cycle, and to see what harmonic structure *it* has. It seems that Bracewell (1953) was the first to recognize this fact. That the spectrum of the raw sunspot data might be unnecessarily complicated compared to a derectified one is very plausible. Consider dividing each cycle of the sunspot numbers into two parts, one half-cycle of low activity, starting half-way down a trailing edge, and the other half of high activity, starting half-way up the next rising edge. These two sets of half-cycles have very different characters: the former have their minima at very nearly the same value of the sunspot number (namely zero), while the latter have maxima varying between about 50 and 200. It is not hard to show that a periodic signal the two half-cycles of which have different shapes must have a spectrum which includes even harmonics.[2]

The quality of the historical data has been much discussed. Eddy (1976)

[2]In this thesis, 'first harmonic' is synonymous with 'fundamental'.

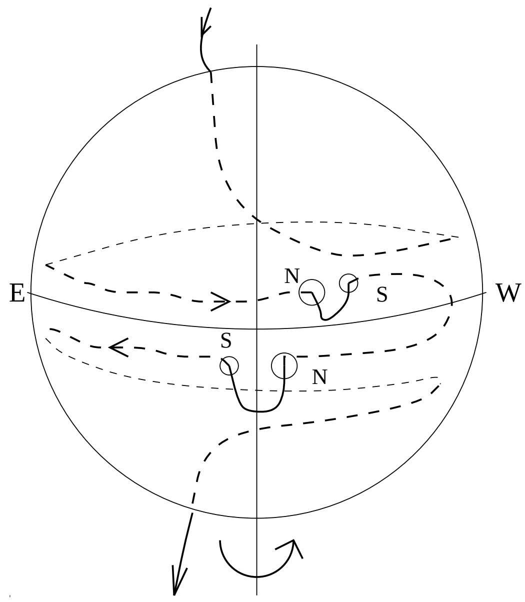


Fig. 5.1. The Babcock model of the solar magnetic field. The differential rotation of the sun acts to twine the initially poloidal magnetic field around inside the sun. The field intensity near the equator increases and eventually flux tubes break through the surface, raised up by magnetic buoyancy. At such places spots form: their polarities are determined from the initial poloidal field. Between cycles the dipole field reverses and hence so do the polarities of the spots of the new cycle.

has called the data after 1818 'good' or 'reliable', but that between 1749 and 1817 'questionable' and that before 1748 'poor'. On the other hand, Dicke (1988) comes to the conclusion that pre-1818 data is probably quite reliable. In this climate of varying and essentially qualitative opinions, it is hard to construct testable information on the errors. However this is not the problem it might seem to be. For there is considerable stability in spectral features constructed from different periods of data, so much so that, for present purposes, a uniform expected error has been assumed, distributed as a Gaussian.

5.2 A toe in the water

The mean annual sunspot index data between 1700 and the present is shown in Figure 5.2. The entropic formalism of Chapter 4 has been used[3] to obtain the most probable amplitude spectrum (Figure 5.3), conditioned on a uniform error of 10 and an 'historically' determined α.[4] The number N of

[3] Brief details of the programs that were used to obtain the results discussed in this chapter can be found in Appendix A.

[4] Since a reliable expected data error is not available, and the most probable spectrum is determined by the combination $\alpha\sigma^2$—both α and σ are required to describe the full

image bins up to the Nyquist limit is 8192 (8K), the number P of data 291. To be noted are: the main peak at a period of 11.05 years, part of at least a quartet of closely equally-spaced frequencies; a lot of low frequency signal; and a group around $\nu = 0.18$ cycles per year (c/y) near where the second harmonic of the main group, which we predicted above, would be expected to appear.

It is interesting to compare this spectrum with the results of Bretthorst's parametric treatment of the same problem (Bretthorst 1988). Not surprisingly the frequencies in his nine term sinusoidal model (*op. cit.* Table 7.1[5]) can be picked out of Figure 5.3. We could not unambiguously say that they were the nine *most* probable frequencies using this spectrum, for the question we are asking is somewhat different. But the connection is apparent. This is a practical demonstration of the symbiosis mentioned in Chapter 2: that a free-form analysis can help in modelling an image, by suggesting parametric forms to describe it. We might call this the *collapse of the hypothesis space.* Of course there is always the possibility that a peak in the spectrum corresponds to something more complicated than a single spectral line. Decay or chirp might be present to account for some of the broadness of the peak. But these features are likely to lead to other clues in the spectrum, such as dispersive line-shapes. And once possible models are suggested, there is a well-defined procedure for fitting and comparing them (Bretthorst 1988). Moral: free-form models are good at suggesting other avenues of approach.

Now to derectification: the essential feature of this is to restore a missing negative sign to alternate cycles. However this is not the whole story. For there is no reason to suppose that the mechanism that rectifies the underlying 'field' does not also nonlinearly modify its magnitude. For example, the sunspot index might reflect the energy in the field, so that a squaring connects field with spots. Bracewell (1988) has made a case for a $\frac{2}{3}$-power derectification. Remarkably, though, the actual form of the nonlinearity turns out to affect to a large extent only the smallest features in the spectrum, the short grass as it were; the forest of main structure is left essentially unaltered. We will use square-root rectified data here.

Figure 5.4 shows this data. Care is needed to identify the sign of some points in the change from one half (magnetic) cycle to the next. In these

posterior—there is no need to be too worried about using the historic recipe. In principle, of course, both α and σ should be considered nuisance parameters which are removed by marginalization. But this is rather an expensive procedure and we are happy to choose σ by eye, based on the cleanness of the best spectrum and the fit to the data. As long as we do not try to fit the data too closely, we are to an extent safeguarded from seeing spurious structure, by the properties of the prior. In fact, the parameter $\alpha\sigma^2$ is like the inverse focal length of zoom lens: as it is decreased we pull structure out of the baseline, until eventually we see noise.

[5] Note typographical errors in the frequencies of lines 7 and 8.

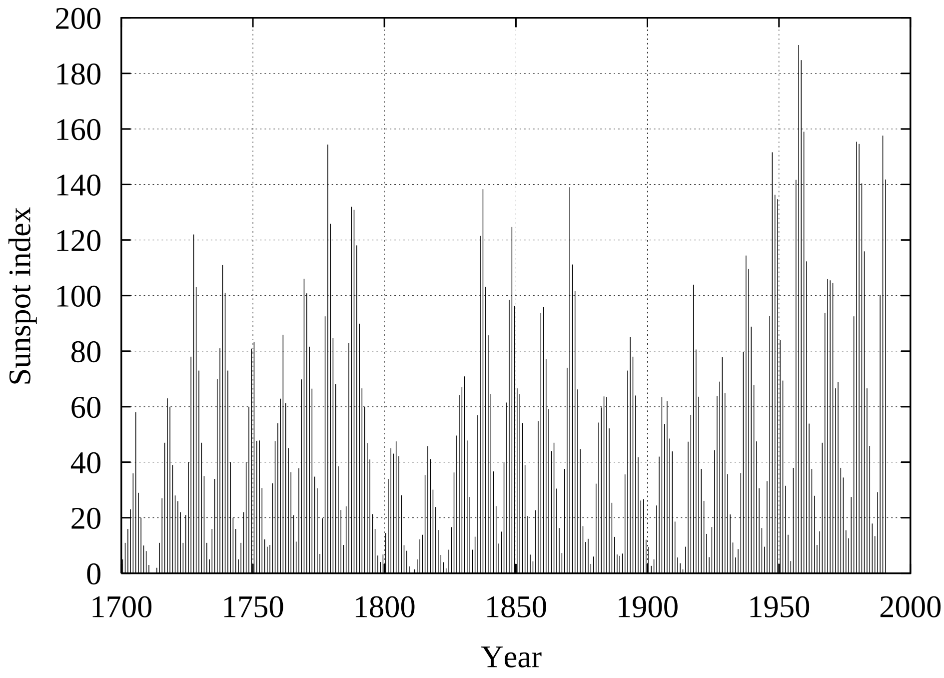


Fig. 5.2. The mean annual sunspot numbers.

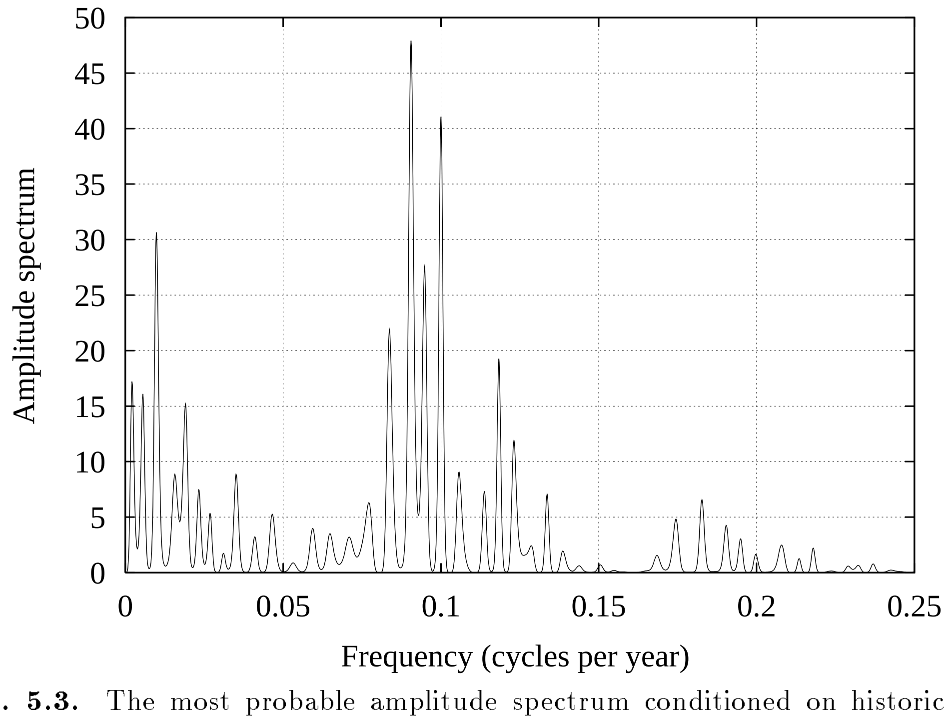


Fig. 5.3. The most probable amplitude spectrum conditioned on historic α, $\sigma = 10$, $N = 8$K and $P = 291$. Only the busy frequency range (up to $\frac{1}{2}\nu_0$) is shown.

Table 5.1. Comparison of amplitude determination via integration and explicit fit to model. Also shown is the overall amplitude of each line derived from columns two and three.

Frequency ν (c/y)	MaxEnt integration H^+	H^-	Parametric fit H^+	H^-	Overall amplitude
0.039	-1.0	$+0.5$	-1.1	$+0.6$	1.1
0.045	-5.7	$+2.2$	-5.6	$+2.1$	6.1
0.050	$+1.5$	-0.2	$+1.3$	$+0.0$	1.5
0.056	-1.1	$+1.3$	-1.0	$+1.4$	1.7

cases, reference to monthly means is helpful. The much more sinusoidal nature of the signal is immediately striking and is confirmed by the greatly simplified amplitude spectrum of Figure 5.5. (The uniform data error is one and $\frac{2}{5}$ of the Nyquist range is shown.) Of course with derectification has come a halving in the frequency of what we call the *fundamental* group, $0.035 < \nu < 0.06\,\mathrm{c/y}$. Note the four almost equally spaced lines that make it up, the second of which is the largest peak and corresponds to the 22.1 year cycle. Figure 5.6 shows the underlying Hartley spectrum.[6] The large amount of low frequency power in Figure 5.3 has been removed. It arose from beating between terms in the underlying field and greatly complicated the spectrum. Lines separated by $\Delta\nu$ in the fundamental group of Figure 5.5 produce a beat at $\Delta\nu$ in Figure 5.3.

The equally-spaced lines in the fundamental suggest at least two models for this group. One is of a carrier that is frequency-modulated, another a (non-sinusoidal) periodic disturbance filtered through a broad resonance.[7]

Given that we interpret the peaks in the spectrum as individual frequencies broadened by truncation effects and, perhaps, by noise, it is easy to determine accurate values for their amplitudes. This is done as described in Chapter 4, by integrating over the peaks in the Hartley spectrum (Figure 5.6). Alternatively we can perform a parametric fit of the amplitudes of the four lines conditioned on their frequencies. With an uninformative prior, this becomes the simple (linear) 'least squares' problem of fitting $s(t) = \sum_i \big(H_i^+ \operatorname{cas}(2\pi\nu_i t) + H_i^- \operatorname{cas}(-2\pi\nu_i t)\big)$ to the data. Amplitudes for the four main peaks in the fundamental region, determined using each of the two methods, are given in Table 5.1. The agreement is reassuringly good. But there are signs in the spectrum that there may be more structure lurking behind them. Note in particular the distorted shapes of the H^- parts of lines 1 and 3. We have not yet resolved everything we might here.

---

[6] The time origin is taken as the year in the middle of the data record, namely 1845.

[7] This is analogous to the voice's production of a vowel sound. In this case the vocal tract passively filters the periodic but non-sinusoidal wave-form produced by the vocal chords.

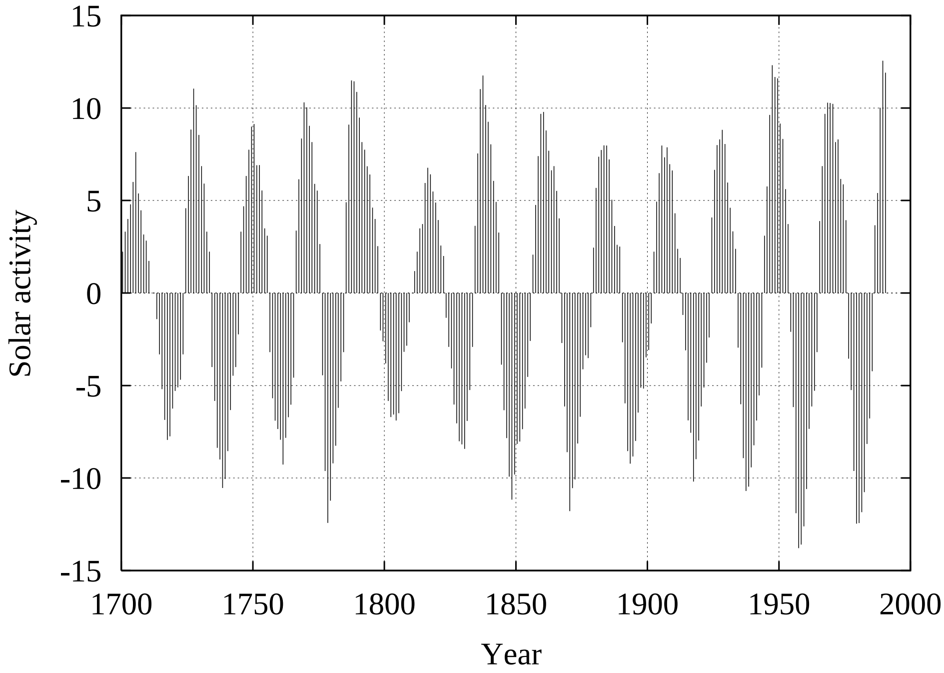


Fig. 5.4. The derectified square-root mean annual sunspot numbers.

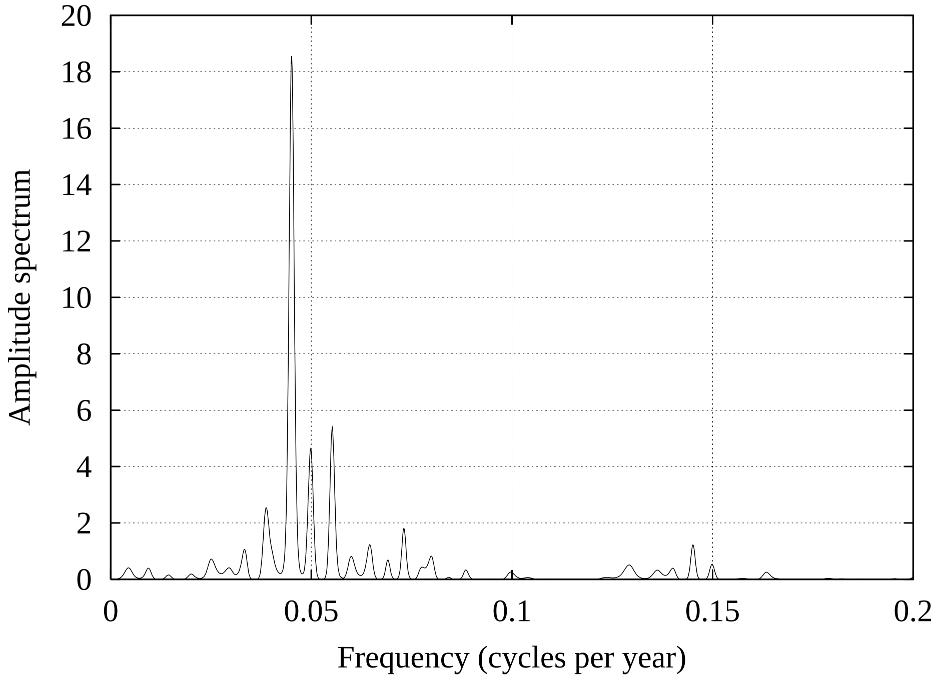


Fig. 5.5. The most probable amplitude spectrum conditioned on historic α, $\sigma = 1$, $N = 8\text{K}$ and $P = 291$. Again only the busy range (up to $\frac{2}{5}\nu_0$) is shown.

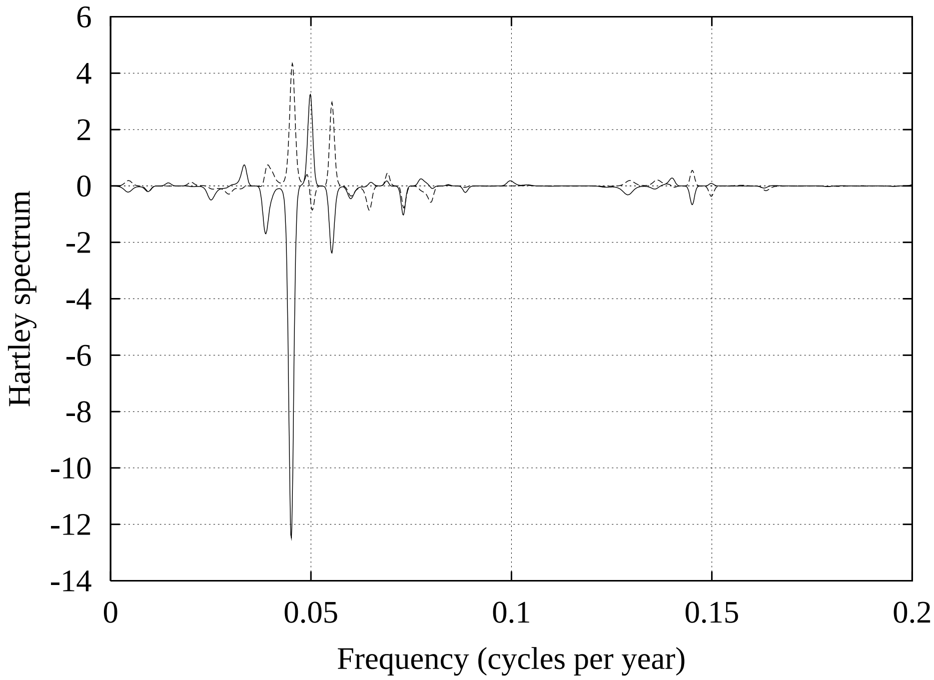


Fig. 5.6. The most probable Hartley spectrum, corresponding to the previous figure. Solid line: $\nu > 0$; dashed line: $\nu < 0$.

Table 5.2. The amplitudes of a five frequency fit to the fundamental.

Frequency ν (c/y)	Hartley components H^+	H^-	Amplitude A
0.040	-1.1	$+0.6$	1.2
0.045	-5.7	$+1.8$	6.0
0.0475	$+0.1$	$+0.8$	0.8
0.050	$+1.4$	-0.4	1.4
0.055	-1.0	$+1.3$	1.6

There is indirect evidence (see §5.5.1) for at least one more line in this fundamental region; the case for a line at $\nu = 0.0475\,\mathrm{c/y}$ is quite good. In fact, this line *can* be resolved in an analysis of the series with the main peak at 0.045 c/y removed. (Because of its size, the 0.045 line tends slightly to distort its neighbours.) It is visible as a shoulder on the 0.05 peak, and is resolved completely if the GSA is left out of the data (§5.5.1). In addition, the 0.039 c/y peak moves closer to 0.04. Table 5.2 summarizes our basic model of the fundamental with the amplitudes determined from a parametric fit.

In Figure 5.7, we show how this five frequency model fits the data. On the whole, it does well. However, around 1780–1820 there is a significant anomaly both in phase and amplitude. We shall see that its spectral

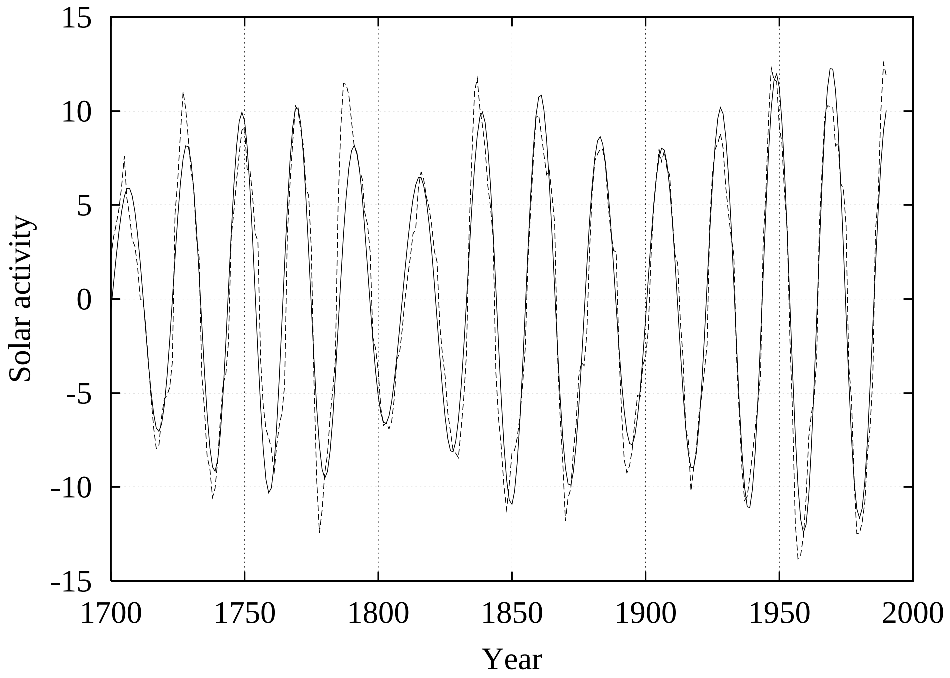


Fig. 5.7. The fit of a five frequency model (solid) to the data (dashed). Note in particular the poor agreement during the period 1780–1820, the so-called GSA.

signature lies elsewhere in the spectrum.

5.3 The third harmonic

We see from Figure 5.5 that there is virtually no power near 0.09 c/y, where a second harmonic might have been expected. This confirms that derectification has produced a signal with no marked differences in the wave-forms of positive and negative half cycles, except, that is, for sign—a considerable simplification. In the region $0.125 < \nu < 0.175$ c/y, there *are* spectral features: it is tempting to identify these as third harmonics. However, we would expect a third harmonic of the main peak to lie near 0.135 c/y, whereas the largest peak in this group is at 0.145 c/y. A slightly different description is required. *A priori*, it is reasonable to expect power here: any error in the non-linear function that was applied as part of the derectification would produce it.[8] However such an error would produce a third harmonic which was *in phase* with its fundamental, in phase in

[8] Suppose that the underlying 'field' is $s(t) = \sin t$ and that its *magnitude* is modified at each instant by a function $f > 0$, so that we observe $f(|s(t)|)$. The sign of s can be inferred as described earlier, but to recover s, we need the function $g \equiv f^{-1}$. If the g which we choose is incorrect, we end up with the signal $\mathrm{sign}(\sin t)g(f(|\sin t|))$. Since gf acts only on the magnitude, and the sign is recovered, this signal has a Fourier series, and hence a spectrum, which contains odd harmonics only; the third harmonic is likely to be the biggest.

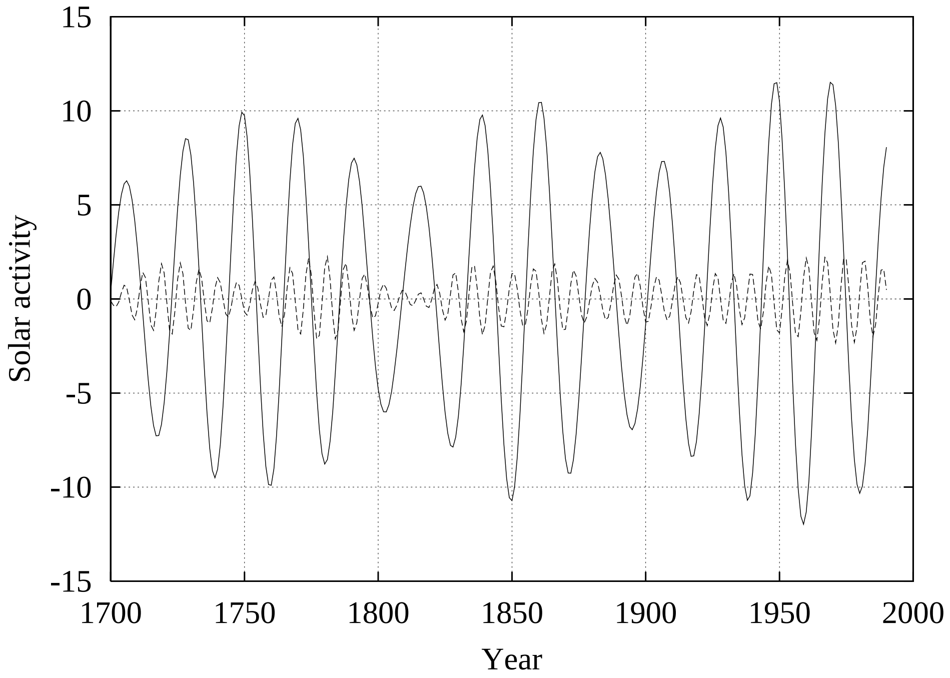


Fig. 5.8. The third harmonic cut (dashed), scaled up by two and plotted with the fundamental cut (solid). A phase lag of the former with respect to the latter of about 1.5 years is clear almost everywhere, as is a correlation in amplitude.

the sense that the zeros of the fundamental would coincide with zeros in the harmonic. This can be checked. Indeed this region of the spectrum is an excellent place for testing the correctness of possible derectification functions.

We can use the spectrum (Figure 5.6) to obtain time-domain representations of particular frequency ranges. This is done by assigning zero to the spectrum outside the range of interest, and applying the Hartley transform to this modified, or filtered, spectrum. Specifically, plots of the signals derived from the fundamental ($0.035 < \nu < 0.058$) and from the group under scrutiny ($0.12 < \nu < 0.17$) will reveal any relationship between the two. Figure 5.8 is a plot of these two time-series. There is a clear phase lag between the two, which is disturbed only during the GSA. Although the period of cycles fluctuates by several years, the third harmonic keeps in quite close step with the fundamental, lagging behind it by a constant 1.5 years. Thus an explanation of this spectral feature solely in terms of an erroneous derectification function is ruled out. We have experimented with powers other than one half in this function but with no significant simplifications emerging in this region of the spectrum. In hindsight it is not hard to see evidence for this feature in the raw data. A significant number of cycles have brief periods of recovery during their decay: observe the crags in Figure 5.4 at 1752, 1850, 1860, 1877, etc., coming on the trailing edges of maxima. Broadly speaking, these appear just where the harmonic cut is at maximum. In addition to the tight phase coherence, there is quite good correlation in the instantaneous amplitudes of the two series.

Bearing this in mind, it is not unreasonable to hope that we might obtain a functional relation between the third harmonic and its fundamental; this would be another significant simplification. The peaking of the harmonic away from 0.135 c/y, makes it hard to believe that we are seeing a strict echoing of the fundamental lines at triple their frequency, especially since the correspondences with the remaining lines is also quite bad. In addition there are many extra interleaved lines in this queer harmonic. Admittedly, the wider spacing there could be allowing extra lines to be resolved: this does not chime well with the evidence. However it is very possible that a function which acted on the combined fundamental, and not on its individual components, could produce these extra lines as cross-terms. And indeed the phase relationships in the main group could result in the greatest power being other than in the peak at 0.135 c/y.

Suppose we wanted to produce a third harmonic from $y(t) = \sin\omega t$. Trigonometrical manipulation quickly reveals that $\sin 3\omega t = 3\omega^{-2}y\dot{y}^2 - y^3$. To delay the harmonic, we can evaluate this expression for some earlier time, in particular 1.5 years earlier. If we replace y by a time domain model of the *whole* fundamental s_1 and add an overall amplitude γ and the time lag τ, we obtain the following model for the third harmonic time-series s_3 (Buck and Macaulay, to appear):

$$s_3(t) = \gamma\left(3\omega^{-2}s_1\dot{s}_1^2 - s_1^3\right)_{t-\tau}. \tag{5.1}$$

This might represent a non-linear memory term in the equations governing the solar cycle. We assign to ω the angular frequency of the main peak. The point about this combination of terms is that to a good approximation it should not put extra power back into the fundamental.[9] This is approximate for two reasons: firstly this is a non-linear function of the fundamental so that superposition does not apply—what was true for the single sinusoid y can only be, at the very best, roughly true for the more complex fundamental; and secondly, a single ω is being employed over the whole fundamental region. Despite these caveats, there is, in practice, remarkably little leakage back into the fundamental.

We demonstrate how well this function performs, in two ways. Firstly, we apply the function s_3 to the filtered fundamental time-series of Figure 5.8 and adjust γ to give a good fit to the third harmonic. In Figure 5.9,

[9] There is another cubic combination of the signal and its derivative that has this property, namely, $\dot{s}^3\omega^{-3} - 3\omega^{-1}\dot{s}s^2$. This can be used, with a different lag, to similar effect. Instead of using an explicit lag, as above, we could employ a linear combination of this term and the unlagged $(s^3,\ \dot{s}^2 s)$ term above. But the results are insensitive to this.

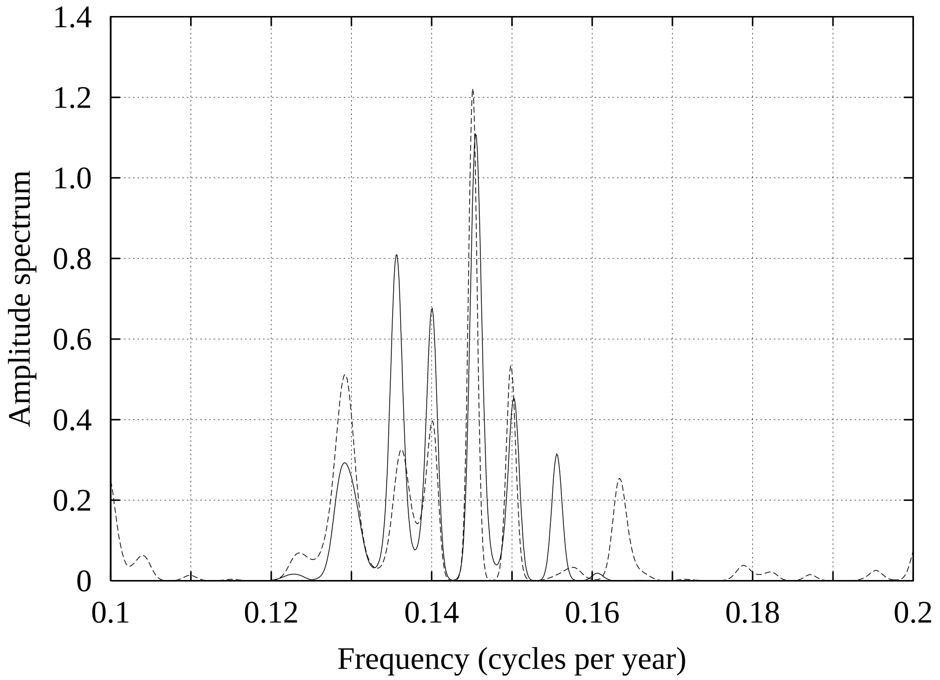


Fig. 5.9. Cubic function applied to the filtered fundamental (solid) shows a qualitative agreement with the spectrum of the data in the third harmonic region (dashed). The time-lag τ is taken as 1.5 years and γ is chosen to fit the overall size. Note, in particular, that the main peak is at 0.145 c/y and not at $3 \times 0.045 = 0.135$ c/y.

we display the spectrum that results, between 0.1 and 0.2 c/y. (As we said above, almost nothing appears below 0.1 c/y.) It agrees semi-quantitatively with the spectrum of the data in the region, which is plotted as well. We have succeeded in producing a model which peaks at 0.145 c/y, but some of the subsidiary peaks are less well reproduced.

The second avenue of approach is more startling and ultimately more informative. We use the time-series of the filtered third harmonic (Figure 5.8) as data, and as a model the cubic function applied to the five frequency model, with frequencies from Table 5.2. The task is to infer the fundamental amplitudes. These turn out to be remarkably close to the amplitudes that appear in the table: for the five lines, they are, in order, 2.1, 5.9, 0.3, 1.8 and 1.2. This is a very stringent test: we have used a tiny perturbation on the signal to infer its macroscopic structure. Figure 5.10 shows the total activity (fundamental plus third harmonic) as determined using this technique. Almost everywhere, the fit is very good. Again, however, it reproduces phase and magnitude rather badly near 1800.

To do any better here, it would be necessary to have some idea of the nature of the non-linearity in the mechanism that drives the solar cycle. What has been shown is that some such effect must be present, and that it

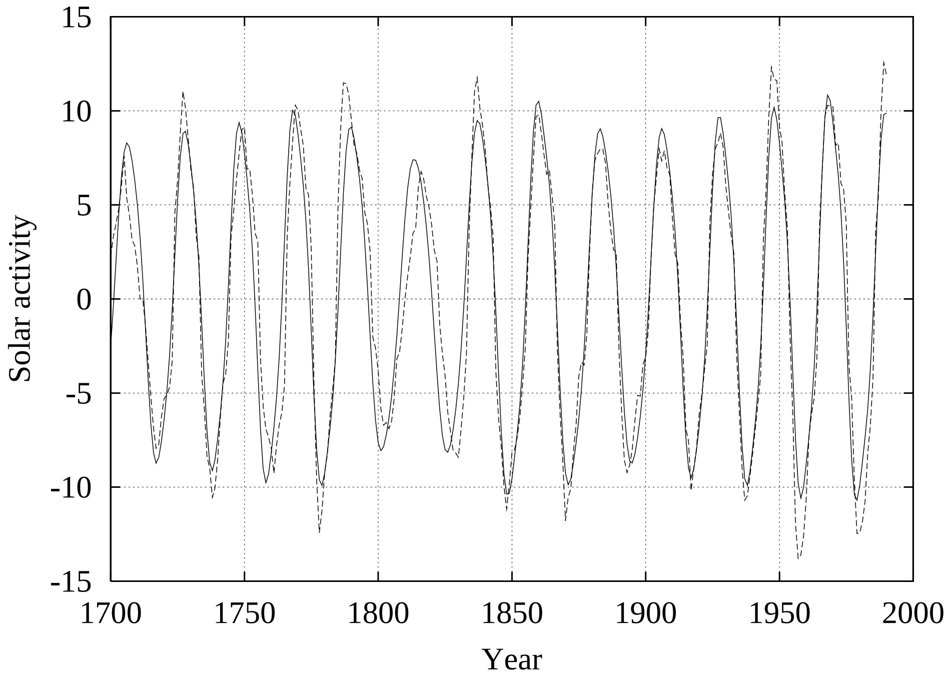


Fig. 5.10. Total activity (fundamental plus harmonic) (solid line) determined from a fit of the 'cubed' five-line fundamental to the third harmonic data only. The data is shown as a dashed line. The agreement is excellent except in the so-called GSA.

can be described remarkably well by a simple cubic function of the signal and its time derivative. The precise details must await a better understanding of the physics of the activity cycle. But there are other clues in the spectrum to non-linear behaviour, as we shall soon see.

On at least two occasions, we have noted that simple models of the fundamental and third harmonic have considerable difficulty fitting a period of the record of about 50 years around 1800. This is the GSA that was mentioned in the introduction. We shall consider it in the next section.

5.4 The Great Solar Anomaly

The data that Figure 5.6, the most probable spectrum, predicts, persistently underfits the measured data just in the region of the GSA, which can only be because the spectrum would need to have considerably more structure to be able to fit the signal here. The budget of entropy decrease versus a better fit to the data just does not balance. To find out where in the spectrum the GSA has its signature, it is helpful to reconstruct a Hartley spectrum which fits the anomaly better. This is done by taking a smaller data error of, say, one half.

It is reasonable to suppose that the fundamental and third harmonics are largely free from the anomaly because they both proved singularly bad

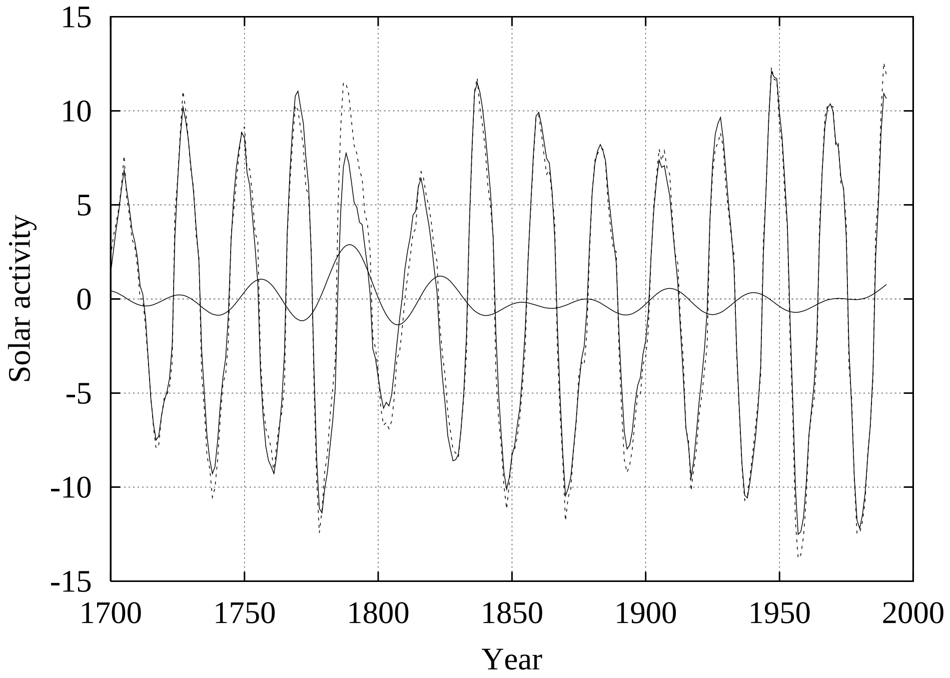


Fig. 5.11. Time-series corresponding to the frequency ranges $0 < \nu < 0.035\,\mathrm{c/y}$ and $0.035 < \nu < 0.5\,\mathrm{c/y}$ (solid lines); the data is shown dashed. The anomaly is very clear around 1775–1810.

at reproducing it. We can investigate the time-domain representations of other regions in the spectrum. The region $0 < \nu < 0.035\,\mathrm{c/y}$ provides strong evidence for the position of the anomaly. We display in Figure 5.11 this series along with the time-series corresponding to the complementary part of the spectrum. Notice the large peak in the low frequency signal near 1790 which reaches an amplitude of three. This is the time of the anomaly, which we see extends from about 1775 to 1810. The spectrum which produced it is notable in that it consists of seven components with frequencies which are multiples of approximately $0.005\,\mathrm{c/y}$. What we are observing here are probably the few first terms in a Fourier *series* expansion of a 200 year period signal, the wave-form of which is localized over a period of no more than 50 years in the GSA. Under this interpretation, the wheel should have turned a complete cycle since 1790 with the result that we are in an anomalous period at the moment.

Now that we have tentatively identified the anomaly, it is interesting to see what a spectrum of the activity cycle looks like when it is excluded from the data. We should expect suppression of the power below $0.035\,\mathrm{c/y}$, but there may be other signs as well. In order to achieve this, we assign the data points in the range 1775–1810 sufficiently large errors that they are effectively excluded and calculate the new maximum entropy spectrum (Figure 5.12).

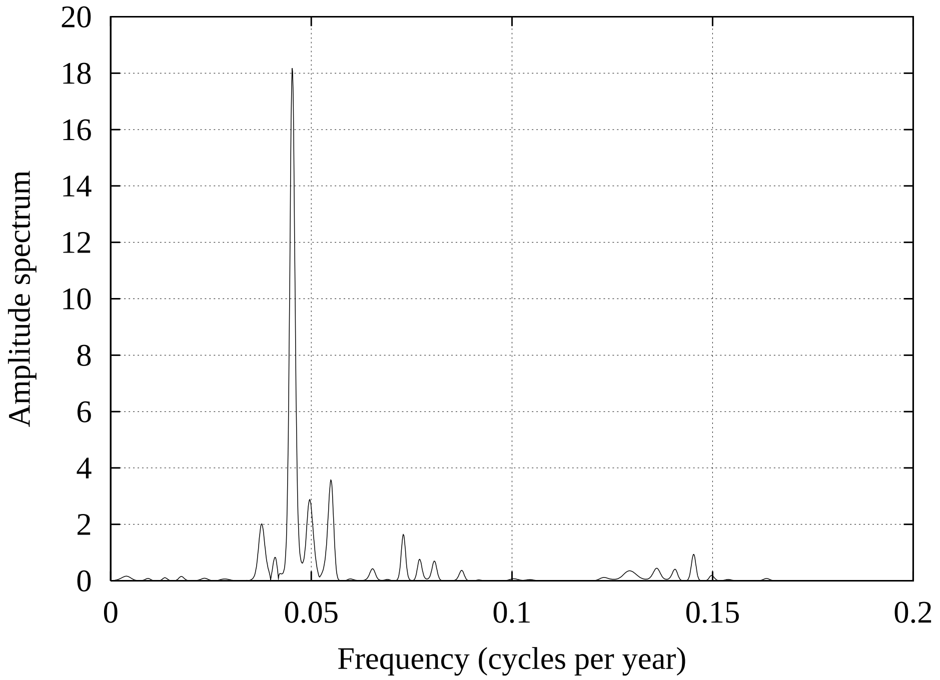


Fig. 5.12. Spectrum of solar activity inferred from a data set which excludes the period of the GSA, 1775–1810. (Actually, data points in this interval are assigned very large expected errors.)

5.5 The Golden Harmonic

5.5.1 The evidence

Figure 5.12 is to be compared with the spectrum for the entire available data (Figure 5.5). The structure below 0.035 c/y has indeed been largely removed. Except for the emergence of a small new peak near 0.04 c/y, both the fundamental and third harmonic are not changed in detail. It is in the region $0.06 < \nu < 0.1$ c/y that a transformation has occurred. Lines have been suppressed and resolution improved—this is clear near 0.08 c/y. The five lines that remain have frequencies 0.065, 0.073, 0.077, 0.081 and 0.088 c/y. A striking relationship to the fundamental's five lines emerges if we calculate the ratios of the frequencies of corresponding lines (Table 5.3); the average ratio is 1.61 (Buck and Macaulay, to appear). The remarkable constancy of this ratio which differs from the golden mean $\phi = (\sqrt{5}+1)/2 \simeq 1.618$ by at most 1.5% cannot be a chance happening, and so we christen this part of the spectrum the **Golden Harmonic**.[10]

Indeed the existence of the 0.0475 c/y peak which is not resolved in Figure 5.5, was inferred from the line at 0.0770 c/y using this model, so that it is a hypothesis with real predictive power. The presence of this

[10] It is not a harmonic in the traditional sense because of its non-integral relation in frequency to the fundamental.

Table 5.3. A comparison of the frequencies of lines in the fundamental with those in the five line group $0.06 < \nu < 0.1$ c/y.

Line number	Frequency (c/y) Fundamental	$0.06 < \nu < 0.1$	Ratio
1	0.040	0.0652	1.63
2	0.045	0.0729	1.62
3	0.0475	0.0770	1.62
4	0.050	0.0807	1.61
5	0.055	0.0875	1.59

line was verified from the greatly improved fit to the data which results from including it in a model, and also from the spectrum of a signal with the main peak subtracted, as mentioned above. We show this spectrum twice, with the GSA first included and then excluded in Figure 5.13. The 0.0475 c/y line shows as a shoulder in the first plot but is fully resolved in the latter, exactly the same pattern as occurs for its golden harmonic shadow. A very small peak near 0.0425 c/y has been pulled out in the GSA-excluded spectrum, although there is no evidence for it in the companion

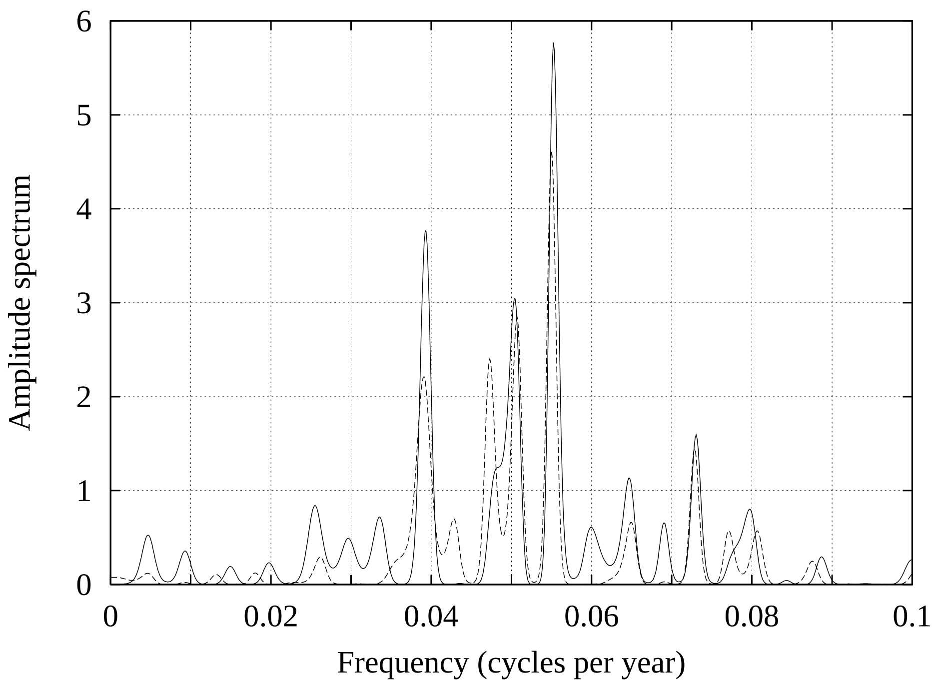


Fig. 5.13. Spectra in the frequency range $0 < \nu < 0.1$ c/y showing the extra structure that is revealed when the main component of the signal, of frequency 0.045 c/y, is subtracted away in the time domain. The solid curve shows the case when the GSA is included in the data, the dashed line the case when it is removed, as in Figure 5.12.

spectrum: we bring in a verdict of *not proven* for this feature. We have now explained the reason for the choice of frequencies in Table 5.2.

5.5.2 Discussion

It is hard to find any linear theory which could generate this golden harmonic out of the fundamental, and so it is necessary to look to the theory of nonlinear systems and the modern theory of chaos. The circle map (Shenker 1982) can generate a spectrum which peaks at frequencies proportional to Fibonacci numbers. These numbers are generated by the recurrence relation $u_n = u_{n-1} + u_{n-2}$ with $u_0 = 1$ and $u_1 = 1$. The ratios of adjacent terms in this sequence quickly approach ϕ. For example, $u_7/u_6 \equiv 21/13 \simeq 1.62$ is already quite a good approximation to ϕ. In this scheme, we might look upon our harmonic as being not perfectly 'golden', but corresponding to two low-lying terms in the Fibonacci series. However the applicability of the circle map to a model of the solar activity cycle is obscure.

The features of the spectrum which have been identified above give clues to the nature of the mechanism of solar activity. The golden and third harmonics are both closely related to the fundamental and are most naturally seen as emerging from small nonlinearities in the equations of motion. Indeed it is possible that the same (approximately cubic) terms which produce the third harmonic could via a transition into quasi-periodicity or chaos also produce the golden harmonic. But that leaves the fundamental to account for. A model for it has previously been proposed (Barnes *et al.* 1980) of white noise filtered through a broad resonance. It seems to me that there is too much regularity in the separation of the individual peaks in this region for the system to be driven by *noise*, and that, in fact, it is a *periodic* signal, of period ~200 years, and localized over a much shorter time, which is filtered through a resonance. In this scheme, the peaks below 0.04 c/y would be the first few Fourier components of the underlying driving signal, with the fundamental corresponding to the small peak near $\nu = 0.005$ c/y.[11] In physical terms, we see the sun suffering a regular magnetic kick, every 200 years, with the field subsequently filtered and nonlinearly shaped as it makes its way to the surface.

5.6 Conclusion

We have used the powerful Bayesian spectral estimation tool developed in the previous chapter to look at the activity of the sun as manifested in the derectified sunspot index. It has allowed us to observe features in the data which have not been discovered before using the cruder techniques of orthodox spectral analysis. In particular, we have seen clear evidence

[11] This fundamental should not be confused with the multiplet of peaks around 0.045 c/y, the fundamental *group*, so called because it contains most of the power in the signal.

for nonlinearity: in the third harmonic which has been shown to have a cubic relation to the fundamental; and in the golden harmonic, a feature which echoes the fundamental's lines at frequencies approximately ϕ times larger, and which we view as a sign of chaos. The fundamental itself is probably best described by resonant filtering of an underlying pulse train with a period close to 200 years.

References

Barnes, J. A., Sargent, H. H. and Tryon, P. V. (1980). Sunspot cycle simulation using random noise. In *The ancient sun: fossil record in earth, moon, and meteorites* (ed. R. O. Pepin, J. A. Eddy and R. B. Merrill), pp. 159–63. Pergamon, New York.

Bracewell, R. N. (1953). The sunspot number series. *Nature*, **171**, 649–50.

Bracewell, R. N. (1988). Three-halves law in sunspot cycle shape. *Monthly Notices of the Royal Astronomical Society*, **230**, 535–50.

Bretthorst, G. L. (1988). *Bayesian spectral analysis and parameter estimation* (Lecture notes in statistics, Vol. 48). Springer-Verlag, New York.

Buck, B. and Macaulay, V. A. (to appear). Fine structure in the sunspot record. *Maximum entropy and Bayesian methods, Paris, 1992.* Proceedings of MaxEnt92.

Coyne, G. V. (1991). Sunspots: the historical background. In *The sun in time* (ed. C. P. Sonett, M. S. Giampapa and M. S. Matthews), pp. 3–10. University of Arizona Press, Tucson.

Dicke, R. H. (1988). The phase variations of the solar cycle. *Solar Physics*, **115**, 171–81.

Eddy, J. A. (1976). The Maunder minimum. *Science*, **192**, 1189–1202.

Foukal, P. V. (1990). *Solar astrophysics.* Wiley, New York.

Shenker, S. J. (1982). Scaling behaviour in a map of a circle onto itself: empirical results. *Physica*, **5D**, 405–11. Reprinted in P. Cvitanović (1984). *Universality in chaos.* Adam Hilger, Bristol.

Sonett, C. P. (1983). The Great Solar Anomaly ca. 1780–1800: an error in compiling the record? *Journal of Geophysical Research*, **88A**, 3225–8.

Waldmeier, M. (1961). *The sunspot-activity in the years 1610–1960.* Schulthess, Zürich.

Conclusion

The analysis of data may not be a glamorous subject of investigation, but when so much devotion and expense goes into collecting data, it must be right that we should make a significant effort to develop the most intellectually rigorous and powerful tools to interpret it. If we are to reason consistently we must use Bayesian probability theory; it is the application of that theory to problems of inversion that has been considered here.

Two problems in physics, one from the arena of inverse scattering, and the other the spectral analysis of a solar time-series, have been studied. They have both necessitated or inspired new developments in the use of entropic priors to reconstruct density functions. For example: a new prior, related to the usual entropic form, has been derived, which can be applied to functions that are not necessarily positive; a technique for determining the reliability of the most probable reconstruction has been described and the resulting error-bars shown to depend on the scale of the features one is trying to infer; it has been shown that the prior for a spectral density function can be a function only of the amplitude of the signal at any frequency and not of the phase; and it has been demonstrated that the use of an entropic prior for the spectral amplitude produces a most probable spectrum that is very clean, with artifacts and noise largely suppressed.

Believable images of nuclear charge densities have been recovered from scattering data, with minimal assumptions about the physics of the nucleus. The entropic model sensibly interpolates and extrapolates the data with realistic reliability estimates.

An entropic spectral analysis of the Wolf sunspot data has been performed and several new features identified, features which are difficult to reconcile with a linear theory of solar activity. These include a time-delayed third harmonic and a group of lines which mimic the fundamental group at frequencies increased by a factor close to the golden ratio. The spectral features of the Great Solar Anomaly have been identified. They are suggestive of a recurring pulse with a period of about 200 years which might be driving the solar cycle.

Further avenues for investigation have been suggested by this work. The error analysis relied on a Gaussian approximation to the posterior probability about its most probable point. This was to avoid the enormously difficult computations of the expected errors, which require integrations over the posterior distribution, a function of a very large number of variables. A closer look at the validity of the approximation and improvements to it would be valuable. New ideas for the estimation of the regularization constant α are needed, which avoid the computational cost of treating it as parameter in the hypothesis space to be marginalized away or the adhockery of $\chi^2 = N$. The method for its determination that was suggested in Chapter 3 goes some way to help in certain problems where the number of 'good' data points can be estimated rather easily, but there is more to be done here. It is my feeling that macroscopic properties of the image and/or the data should allow reasonable values for this constant to be determined.

There are numerous possible applications of the work described here. They include detailed analyses of the many kinds of time-series that crop up in various fields, from dendrochronology to dynamical systems, speech processing to astrophysics. Another example is plasma diagnostics which requires just the sorts of tools that have been developed here to analyse time-series and to make inferences from scattering data. Indeed, the use of entropic priors would bring benefits to all attempts to invert scattering data and the quantification that the theory brings would allow much more realistic determinations than have been possible up to now of the kinds of structures that can be reliably inferred from limited data.

Appendix A: Implementation

The reconstructions of Chapter 3 were determined from two computer programs, the first of which sought the most probable charge distribution, while the second constructed best estimates and error bars on interesting linear functionals of that distribution, specifically the total amount of charge in certain extended domains and the predicted form factor at various values of the momentum transfer. Figure A.1 shows the flow of information through these two modules, respectively A and B.

The most probable image $\boldsymbol{f}_0$ is easily seen to be the one for which $Q(\boldsymbol{f}) \equiv \alpha S(\boldsymbol{f}) + L(\boldsymbol{f})$ is a maximum.[1] This image is clearly an implicit function of α, the regularizing constant. The modified discrepancy principle described in Section 3.4 is used to obtain a plausible value for α. That is, we need to maximize Q with respect to $\boldsymbol{f}$ and adjust α so that the constraint $\chi^2\big(\boldsymbol{f}_0(\alpha)\big) = P - W$ holds. This job is performed by the first program (A). The algorithm used is based on a stripped down version by Daniell (private communication) of the 'Cambridge algorithm' of Skilling and Bryan (1984) (also Skilling 1986). This is a search algorithm, the form of which is shown in the flow diagram of Figure A.2. I shall now briefly describe how it works.

Optimization algorithms such as the conjugate gradients method seek the maximum of a function by successive line searches along directions chosen in some sensible way. Their success relies on the function concerned being reliably quadratic over many line searches, so that information on the curvature (second derivative) matrix can be built up cumulatively. Unfortunately the entropy function is highly non-quadratic,[2] that is, a quadratic approximation at any image point is valid only over a very small region. Thus the present algorithm does not try to carry forward information on the curvature between iterations. Instead, at each stage, a fresh quadratic approximation to Q is computed, but, in order to produce a powerful enough

[1] Here the slowly varying measure function is neglected.

[2] This is why it is good at suppressing noise and truncation artifacts, as we saw in Chapter 2.

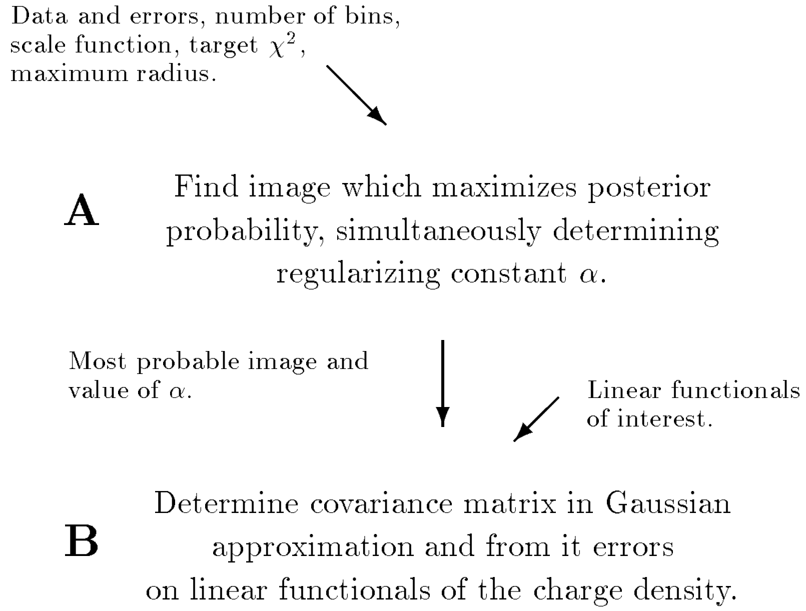


Fig. A.1. The flow of information through the two programs that were used to produce the images of Chapter 3.

algorithm, that approximation is not along just a *single* search direction but rather it is in a larger space spanned by a *set* of plausible directions. Three directions have been found to be adequate in many cases. They are taken to be along ∇S, ∇L and $\nabla\nabla L \cdot (\widehat{\nabla S} - \widehat{\nabla L})$, all evaluated at the current image point; the hats denote unit vectors.[3]

In order to make it easier for large amplitudes to build up quickly in the image where they are required, another trick is used. A metric tensor is introduced into all scalar products between image vectors. That tensor is required to boost the components of search vectors where the image is large. A very natural[4] metric g^{ij} is $-[\nabla\nabla S]^{-1}_{ij}$, which equals $\delta_{ij}\big({f^i}^2 + {m^i}^2\big)^{\frac{1}{2}}$ (no summation). Then, any derivative with respect to the (contravariant) image f^i can only be used to update the image if it too is contravariant, that is, if it has been multiplied by g^{ij}. This modifies the search directions as required. It helps in another way too. The quadratic approximation is only good within some distance of the current image point. If the increment to $\boldsymbol{f}$, namely $\delta\boldsymbol{f}$, is too big, it should be reduced. If we define the size of $\delta\boldsymbol{f}$ using the metric: $|\delta\boldsymbol{f}|^2 = g_{ij}\delta\boldsymbol{f}^i\delta\boldsymbol{f}^j = \sum_i(\delta\boldsymbol{f}^i)^2/\big({f^i}^2 + {m^i}^2\big)^{\frac{1}{2}}$, it is

[3] The first two directions are natural choices because they both appear in the gradient of Q. The third comes from considering the approximate value of ∇L, the second search direction, after a single iteration.

[4] There are many indications (e.g., Rodriguez 1989; Skilling 1991) that the 'entropy metric' may be very fundamental to our fuller understanding of the status of priors for images.

1. Input data, errors, target χ^2,
scale function m, no. of image bins, radius.

2. Initialize image $\boldsymbol{f}$ to zero.

3. Construct the three search directions
at current $\boldsymbol{f}$: $\nabla^i S$, $\nabla^i L$, $\nabla^i \nabla^j L\,(\widehat{\nabla_j S} - \widehat{\nabla_j L})$.

4. Converged? That is,
are ∇L and ∇S closely parallel?
is χ^2 close enough to its target value?

YES → 5. Output $\boldsymbol{f}_0$ and
predicted data, $\hat{O}_p \boldsymbol{f}_0$.

NO ↓

6. Construct quadratic approximation
to S and L in subspace.

7. Determine the χ^2 to aim for on this iteration
and the appropriate α.

8. Determine change in $\boldsymbol{f}$ that maximizes
$\alpha S + L$, in quadratic approximation.

9. Reduce change $\delta \boldsymbol{f}$
if it is too ambitious.

10. Update image $\boldsymbol{f} \rightarrow \boldsymbol{f} + \delta \boldsymbol{f}$ and
output diagnostics.

Fig. A.2. A flow diagram of a procedure for maximizing $\alpha S + L$ and simultaneously determining a good value for α. See the text for explanation.

clear that larger increments will be allowed when the image itself is large. In this way the image will not be prevented from developing structure. It should be kept in mind that introducing a non-trivial metric is merely a computational device: the equations to be solved are unchanged.

With all the relevant data and parameters input in Stage 1 and the image initialized in Stage 2, it is the construction of the (contravariant) search directions which occurs at Stage 3. This requires a number of applications of the image space to data space transformation (`OPUS`) and its transpose (`TROPUS`), in order to calculate the gradient and Hessian of L. For the charge density problem, where the sizes of both spaces were quite modest, simple matrix multiplications were used in these functions. However, in the much larger spectral inversions of Chapter 5, Fast Fourier Transforms were employed for both `OPUS` and `TROPUS`, in order to reap the benefits of $N \ln_2 N$ complexity.

From the first two search directions, a crucial diagnostic parameter can be determined which tells us whether the image which maximizes Q has been found yet (Stage 4). For, at the solution, $\nabla Q = 0$ and hence ∇S is parallel to $-\nabla L$. Since there can be only a single maximum, the angle between these two directions determines whether the algorithm has converged: when this angle is less than some threshold, the maximum can effectively be considered reached. We must, of course, also check that $\chi^2(\boldsymbol{f})$ is close enough to the target value supplied at Stage 1. If it is, we stop (Stage 5); otherwise, we continue (Stage 6).

Because the Hessian of $-Q$ is positive-definite, the quadratic approximation to Q has a single maximum in the search space, for all $\alpha > 0$. It is easily determined. We write

$$S(\boldsymbol{f}+\delta\boldsymbol{f}) = S(\boldsymbol{f}) + \sum_{\mu=1}^{3} x^\mu s'_\mu + \tfrac{1}{2} \sum_{\mu,\nu=1}^{3} x^\mu x^\nu s''_{\mu\nu}, \tag{A.2}$$

$$L(\boldsymbol{f}+\delta\boldsymbol{f}) = L(\boldsymbol{f}) + \sum_{\mu=1}^{3} x^\mu l'_\mu + \tfrac{1}{2} \sum_{\mu,\nu=1}^{3} x^\mu x^\nu l''_{\mu\nu}, \tag{A.3}$$

where $\delta\boldsymbol{f} = \sum_\mu x^\mu \boldsymbol{e}_\mu$ is the displacement in the sub-space. The three $\boldsymbol{e}_\mu$ are unit contravariant vectors in the search directions. For example, $\boldsymbol{e}_1^i = g^{ij}\nabla_j S/\sqrt{\nabla^i S\,\nabla_i S}$.[5] The three s'_μ are the gradients of S in the three search directions: $s'_\mu = \boldsymbol{e}_\mu \cdot \nabla S$, and the 3×3 matrix $s''_{\mu\nu}$ is the Hessian of S in that space: $s''_{\mu\nu} = \boldsymbol{e}_\mu \cdot \nabla\nabla S \cdot \boldsymbol{e}_\nu$. Similarly, the quantities l' and l'' are derivatives of L in the sub-space. The stationarity condition $\nabla Q = 0$ becomes

$$(\alpha s''_{\mu\nu} + l''_{\mu\nu})x^\nu = -(\alpha s'_\mu + l'_\mu). \tag{A.4}$$

[5] The operator ∇_i is $\partial/\partial f^i$, the position of the index implying a covariant vector; its contravariant sibling is $\nabla^i = g^{ij}\nabla_j$.

This is a set of three linear equations in the three unknowns x^μ, which is solved by standard techniques (Press *et al.* 1986, Chapter 2).

But all this is based on knowing what value α has. In fact we must determine this constant as well, during the iteration. Fortunately, we can determine the x^μ that solves (A.4) for any α, and hence, by substituting it into (A.3), the $\chi^2(=-2L)$ that it implies. In addition, the lowest value that χ^2 can have in the current sub-space, $\chi^2_{\rm low}$, occurs when there is no regularization, that is, when $\alpha = 0$. As α increases, χ^2 increases as well. If $\chi^2_{\rm low}$ is less than the target χ^2, $\chi^2_{\rm target}$, we choose to to aim for $\chi^2_{\rm target}$ on this iteration: this is just the value we want. Otherwise, a less ambitious value half-way between the present χ^2 and $\chi^2_{\rm low}$ is taken: this ensures that δQ has a contribution coming from the entropy S. Now that a suitable value of χ^2 to aim for on this iteration has been determined, the α that produces it is sought. A crude search which traps α between progressively closer and closer lower and upper bounds allows α to be found (Stage 7). Then, (A.4) is solved for x^ν (Stage 8) and hence $\delta \boldsymbol{f}$ is determined. Its squared magnitude $|\delta \boldsymbol{f}|^2$ is calculated and if it is too big for the quadratic approximation to be good (greater than $0.1 \times \sum_i |f_i|$, say), the change in the image $\delta \boldsymbol{f}$ is multiplied by a suitable constant less than one to reduce its magnitude (Stage 9). Finally the image is updated and diagnostics such as the current χ^2 and the angle between ∇S and $-\nabla L$ displayed. The next iteration starts back at Stage 3. This iterative scheme is robust enough for many linear problems: the examples in this thesis were calculated using it.

Slight changes were necessary for the spectral problems of Chapter 5. Of course, the image-to-data transformation is different. It is a Hartley transform and, as I said above, it was calculated using a Fast Fourier transform, which provides a useful economy for images of 2^{13} bins. In addition, the entropy is a more complicated function, since it is no longer a simple sum of terms each a function only of the image in a single bin. In fact, as we saw in Chapter 4, it is a sum of terms which are functions of the pairs of amplitudes $H(\nu_i)$ and $H(-\nu_i)$. This means that derivatives of S are rather more complicated. Finally, the metric employed, although related to the Hessian of the entropy in the same way as before, is not diagonal, which makes 'raising an index' slightly more difficult.

In Chapter 3, it was shown how to calculate error estimates on the recovered densities and how to estimate various linear functionals. A program (B) evaluates these quantities. At its heart is a routine to calculate the covariance matrix $\mathbf{C}$ of the posterior distribution in Gaussian approximation. From this matrix and the most probable image, all such quantities can be calculated using (3.27) and (3.28). The inverse of $\mathbf{C}$ is directly accessible from (3.23), all the terms being at hand: $\mathbf{g}$ is the square root of the metric tensor g^{ij}, a good α is available from (A) and $\nabla\nabla L$ is $-\sum_p O_{pi} O_{pj}/\sigma_p^2$. However even for an image of $N = 256$ pixels, numerical inversion of $\mathbf{C}^{-1}$ using standard routines (Press *et al.* 1986, Chapter 2) would be expensive, having N^3 complexity. In fact, this is not necessary, as we showed in Section 3.6. Because of the special form of $\mathbf{C}^{-1}$, it suffices to determine the non-zero eigenvalues and the corresponding eigenvectors of $-\mathbf{g}^{-1}\nabla\nabla L\,\mathbf{g}^{-1}$. Then, (3.26) can be used to obtain $\mathbf{C}$. In the same section, I showed how these quantities can be determined from diagonalizing a much smaller $P \times P$ matrix, by exploiting both the form of $\nabla\nabla L$ and the singular value decomposition (see the appendix to Macaulay and Buck 1989 for a discussion of the SVD). This smaller matrix is diagonalized using Jacobi rotation (Press *et al.* 1986, Chapter 11), and from its eigenvalues and eigenvectors, the covariance matrix is constructed using (3.26) and the expression for $\mathbf{U}$ given below that equation, by a handful of matrix multiplications.[6]

The linear functionals that we want to infer are input at the start of the program in the form of the appropriate sets of a_is: see the paragraph containing (3.27). If we want an estimate of the amount of charge in bin j, then each a_i equals δ_{ij}. From (3.27), the most probable value is then just the value that was calculated for that bin in program (A), and, from (3.28), the squared error is the (i,i)th element on the diagonal of $\mathbf{C}$. In Figures 3.6 and 3.8, the inferences are the charge densities in variously sized macrobins, that is, neighbouring bins grouped together. In this case, for any macrobin, $a_i = \frac{1}{\delta}$, if i is a bin in the macrobin and zero otherwise; δ is the macrobin's width. Again (3.27) and (3.28) are used to determine the most probable value and the expected error. Notice that more elements of the covariance matrix are used, since the set of a_is is more complicated. And then in Figures 3.7 and 3.9, for which the predicted form factor at various values of q was calculated, the a_is took the values $j_0(qr_i)$, for each q required. For this yet more complicated set of a_is, the whole of $\mathbf{C}$ was needed for each inference.

Finally, and turning again to Chapter 5, the least-squares examples there, such as those of Tables 5.1 and 5.2, in which the model of the signal was a linear function of the parameters, and that of Figure 5.10, in which there were non-linear parameters, were computed using the Levenburg–Marquardt method (Press *et al.* 1986, Section 14.4), another algorithm to find an extremum.

References

Macaulay, V. A. and Buck, B. (1989). Linear inversion by the method of maximum entropy. *Inverse Problems*, **5**, 859–73.

---

[6] If further speed improvements are required, it should be noted that the only eigenvectors of the $P \times P$ matrix that are needed are those with substantial eigenvalues. This can surely be exploited to reduce the computational burden.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1986). *Numerical recipes: the art of scientific computing.* Cambridge University Press.

Rodriguez, C. C. (1989). The metrics induced by the Kullback number. In *Maximum entropy and Bayesian methods, Cambridge, England, 1988* (ed. J. Skilling), pp. 415–22. Kluwer, Dordrecht.

Skilling, J. (1986). The Cambridge maximum entropy algorithm. In *Maximum entropy and Bayesian methods in applied statistics: proceedings of the fourth maximum entropy workshop, University of Calgary, 1984* (ed. J. H. Justice), pp. 179–93. Cambridge University Press.

Skilling, J. (1991). Fundamentals of MaxEnt in data analysis. In *Maximum entropy in action: a collection of expository essays* (ed. B. Buck and V. A. Macaulay), pp. 19–40. Clarendon Press, Oxford.

Skilling, J. and Bryan, R. K. (1984). Maximum entropy image reconstruction: general algorithm. *Monthly Notices of the Royal Astronomical Society*, **211**, 111–24.