Irregularly Sampled Signals: Theories and Techniques for Analysis

Richard James Martin

GEC-Marconi Ltd.
University College London

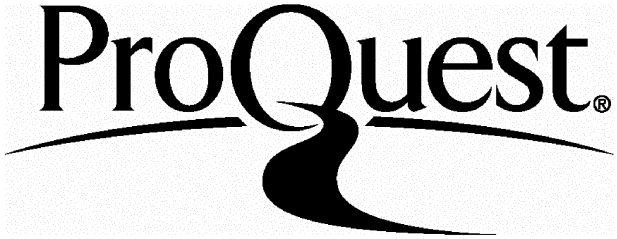
This thesis is submitted for the degree Doctor of Philosophy
January 1998

ProQuest Number: U642081

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest U642081

Published by ProQuest LLC(2015). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code.
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

To Michael Kearney
who introduced me to time series analysis and much more besides

Abstract

This thesis is about sampling theory and methods for analysing signals that have been sampled at irregularly spaced points.

Irregular sampling may arise naturally (examples of its occurrence may be found in geophysics, tomography, astronomy, and laser anemometry). In many cases it presents difficulties because standard techniques are unable to cope with the uneven sampling. However there is an alternative and exciting facet to the subject: deliberate aperiodic sampling. This is being mooted as a method for unambiguous frequency identification in new generations of signal analysers and of pulse-Doppler and synthetic-aperture radars. For the classes of signal that these systems need to process and analyse, signal reconstruction is not of prime importance and it can be the wrong approach. The principal aim of this thesis is to develop methods for analysing irregularly sampled data and the principal theme is methods that do not employ explicit signal reconstruction.

The key contributions of this thesis are the development of prediction and filtering. A difficult problem associated with the spectral analysis of irregularly sampled signals is that the dynamic range of the observed spectrum is greatly reduced. It can however be resolved using a combination of elementary spectral analysis and advanced linear filtering techniques. The fast optimal filtering algorithms enable this to be done. They are derived using our general theory of linear prediction, which we extensively test on synthetic data.

Other important contributions are made in the theories of nonlinear prediction and of sampling series. Nonlinear techniques are designed for signals of dynamical origin and we show that they can be made to work for irregular sampling. The work on sampling series shows that classical signal processing techniques such as system identification, convolution and filtering are not the preserve of regular sampling.

Additionally an extensive review of sampling theory and its relation to signal processing is included. It provides an in-depth introduction to the subject and its fascinating literature.

Acknowledgements

Any thesis, particularly one on a broad subject, is likely to have been influenced by many people, and that is the case here. I have received much encouragement and sound guidance from my supervisor, Dr. Jaroslav Stark (UCL), and benefited from wise advice from my secondary supervisor, Dr. Stephen Bishop (UCL).

I have been fortunate to pursue this research while employed by GEC-Marconi Ltd.; irregular sampling, as a project, was suggested and supported by Dr. Bill Bardo (GEC-Marconi HQ, Stanmore), and as a PhD topic by Prof. Michael Kearney (formerly of Hirst Research Centre). I appreciate the support that I have received from people in my laboratory, in particular from Dr. Ayman El-Fatatry.

In 1995 I was awarded an Industrial Fellowship by the Royal Commission for the Exhibition of 1851. These Fellowships enable industrial scientists to pursue academic study and research and I am most grateful to the Commissioners for this opportunity. The generous travel grant has enabled me to work with researchers whom I would otherwise not have met. I am grateful to Prof. Abdul Jerri (Clarkson University, NY) and Prof. Hans Feichtinger (Vienna University) for their hospitality and for stimulating discussions, and am particularly indebted to Prof. Ken Lever of the Signal Processing Research Institute, Adelaide. Prof. Lever, formerly of Hirst Research Centre, was a generous host during my two extended visits to Australia and his keen interest in promoting sampling theory and nonuniform sampling methods has thrown up many research opportunities.

This work has also benefited from discussions with Dr. Simon Godsill (Cambridge University), Prof. Rowland Higgins (Anglia University), Prof. Gerry Cain (University of Westminster) and Mr. Michael Clayton (formerly of Hirst Research Centre), after which various infelicities were removed. It is hard to see how the work of §3.3 would have come about had not Dr. Chris Brookes, my Director of Studies at Cambridge in times gone by, pointed me to a book on exponential Diophantine equations, when by chance I mentioned that I was in trouble proving a certain conjecture on sums of sinusoids.

A final 'thank you' must go to Dr. David Castelow and Dr. Neil Cade, both of GEC-Marconi Research Centre, for their kind offers to read the manuscript, and to Mrs. Rita Brown (formerly of the Hirst Research Centre Library), who made my literature survey a much easier task. David's mastery of LaTeX and our Sun/UNIX system made smooth the preparation of the thesis and got me out of a fair few scrapes. Nonetheless the responsibility for any errors in the thesis rests entirely with me.

Contents

List of Figures

Chapter 1

Introduction

1.1 Sampling theory and irregular sampling

Man's view of the world around him is formed from observations or *samples*. In observing a static object one may make observations at different points in space. If the object moves about, observations may be made on its position, periodically or aperiodically with time. For example in navigation one might be observing someone else's ship from land, or calculating the position of one's own ship relative to a fixed frame of reference. More generally the physical properties or qualities of an object (temperature, size, colour, etc.) may vary with time. In addition we may wish to observe an object that is abstract rather than concrete, such as a radar or radio waveform, or a sound wave; in these cases it is the electromagnetic field or the density of the medium that is varying with time.

We shall be talking about signals; a signal is any time-varying quantity, and signals may be classified in a variety of ways (periodic, stochastic, deterministic, linear, nonlinear, and so on; see e.g. [135]). At some juncture one must ask how samples of a continuously-varying quantity relate to that continuously-varying quantity. This question is answered by signal sampling theory (or *sampling theory* for short). One might also ask what a signal (continuous or discrete) says about the system from which it came. That is the science and art of *signal analysis*.

Sampling theory has a long history and, in common with many other areas of mathematics, finds its roots in the work of Cauchy [46] and Gauss. Its name has almost become synonymous with that of C. E. Shannon, who amongst others is credited with the statement of the 'sampling theorem'. This states that from periodic observations one may reconstruct a signal that contains no frequency above half the sampling rate (a limit to which Nyquist's name has become attached). In other words, for a correct representation one must sample at least twice per cycle of the highest frequency component. The importance and attraction of this theorem are that the samples of a bandlimited signal contain all the information needed to reconstruct that signal. But like all good theorems it raised far more questions than it answered, as it was not long before people began to ask whether the sampling has to be regular, and what to do when it is not; or how to generalise the notion of frequency, or to drop the requirement of 'bandlimitedness'.

The theory has also been given practical impetus from situations in which irregular sampling arises naturally [143]§VIII. Some examples and pointers to literature are:

- Geophysics [62, 168, 169, 183]. These problems are usually spatial rather than temporal. One wishes to collect geophysical data (e.g. electrical resistivity of the ground, gravitational or magnetic potentials). Such scalar or vector fields can only be sampled at points on the Earth's surface (or down holes) to which the investigator has access. The observations therefore tend to be irregular, and clustered, because on setting up the equipment the user finds it convenient to make several sets of readings close by.

- Computer tomography [197]. 'Tomography' means 'picture-cutting'. A 3D object (such as a human head) is to be analysed from 2D pictures. Each 2D picture is obtained by passing an X-ray or positron beam through the object and recording the transmitted image; the procedure is repeated for different observation angles. As it is generally convenient to scan in spirals, and as the geometry is polar rather than Cartesian, the observations are not on a simple Cartesian grid.

- Astronomy [180]. When observing a star, for example, one only has access to it at certain times of the day or the year (owing to the Earth's rotation and orbiting). There are also the difficulties of adverse weather conditions preventing observations being made, and of equipment faults.

- Laser anemometry [10]. The objective is to 'seed' a gas flow with small particles (in turbomachinery the particle size is less than a micron or so) and illuminate with laser light (see e.g. [131] for an overview). In laser Doppler anemometry a pair of crossed laser beams intersect at a spot and interfere to produce fringes. As a particle (following the gas flow field) crosses these fringes transversely a sequence of flashes is observed by an optical detector. Particles turn up at irregular intervals, so one has an irregularly sampled record of a time-varying flow field.

- 'Spiky data' e.g. heart beat [58, 59, 176, 177],[106]§9. In recent years there has been increasing interest in understanding the pattern of the heart beat from the point of view of specrtral analysis or of nonlinear dynamics. Electronic oscillators [218, 50, 112, 130] may also give spiky outputs. In each case the interspike spacing may be taken as the observed variable. Work on spectral analysis uses the sequence of spacings as an irregularly sampled time series (if (t_n) are the times at which spikes occur, the ordinate is $\frac{1}{2}(t_{n+1} - t_n)$ and the abscissa is $\frac{1}{2}(t_n + t_{n+1})$). More generally one can observe for a multivariable system the state variable Y whenever the variable X performs a spike, so that the sequence of Y-observations is in a very natural way an irregularly sampled data set.

- Control theory [143]§VIIIA. The objective is to sample adaptively and irregularly to reduce the volume of data that the controller has to cope with. One

can imagine that this might work well if the controller needs to be active only when the system enters a certain region of its state space.

- Missing data [65, 66, 67, 68]. Any data stream subject to 'drop-outs' gives rise to a missing data problem. An obvious example is audio restoration, to which [74] is a fine introduction. The subject of missing data includes the additional question of how to detect corrupted samples.

In the first three of these applications it is probably fair to say that irregular sampling is undesirable. This is because one wants to reconstruct the continuous-time waveform, which is assumed to be slowly-varying. When the sampling is irregular the sampling theorem no longer applies and reconstruction has to use methods that are computationally intensive. Signal processing tools such as the Fast Fourier Transform (FFT)—used for spectral estimation, convolution and filtering—no longer work. It is not just that there is no *fast* discrete Fourier transform (DFT); it is that the irregular DFT is not even invertible, so one cannot go back and forth between time and frequency domains. That is what make convolution and filtering so difficult.

Summers [209], in his PhD thesis, poses two questions about irregular sampling:

- What are the disadvantages of irregular sampling (why is it not used very often in sampled-data systems)?
- What benefits can be gained from irregular sampling?

We have answered the first question in much the same way that Summers does, though he seems less clear about the consequences of noninvertibility of the DFT.

Let us now turn to the second question, to which the answers have a more positive flavour. The key point is that by regular sampling one cannot unambiguously identify frequencies above the Nyquist critical frequency of half the sampling rate. Signals above that critical frequency get folded back into the interval $[-\frac{1}{2}f_s, \frac{1}{2}f_s]$ (f_s is the sampling rate) and be indistinguishable from those that do in fact have frequencies in that interval; this effect is known as *aliasing*. With irregular sampling this restriction disappears. Irregular sampling is therefore described as an anti-aliasing measure. This has led to two major applications:

- High-speed signal analysers. The best example is the 'digital alias-free signal processing system' (DASP) pioneered by groups at the Institute of Electronics and Computer Science in Riga (Latvia) and the University of Westminster (London); see [20, 21] for an overview. It is currently able to identify components at frequencies up to 1.2GHz, far higher than the maximum rate at which analogue-to-digital converters (ADCs) may be operated using today's technology—yet its ADC samples at an average rate of only 80MHz. The Nyquist limit has been exceeded by a factor of 30. This technology presents formidable hardware problems, in addition to the difficulties of processing the signal samples. The most serious obstacle in the hardware has been the controlling and recording of the sampling instants to an accuracy of a few tens

of picoseconds; time-jitter errors seriously reduce the accuracy of frequency estimation.

- Radar [14] and synthetic-aperture radar [120].

  Consider a pulse-Doppler radar, which uses the pulse round-trip time to estimate range, and the Doppler shift to estimate radial velocity. The pulse repetition frequency (PRF) is an important parameter to be set [109]. If the range information is to be unambiguous for targets up to 150km, a PRF less than 1kHz is required. But for unambiguous identification of the Doppler frequency a PRF of at least ≈100kHz is required (this is because fighter aircraft radars at X-band (≈10GHz) typically encounter Doppler shifts of 60–120kHz). These constraints are mutually contradictory. One solution is to use medium PRF (MPRF) schemes, in which multiple PRFs are transmitted [109]; irregular pulses are also being tried as an opportunity to reduce aliasing in the range and Doppler domains [14].

  Synthetic aperture radar[1] or SAR is a well-established technique for imaging the ground to one side of an airborne platform. Moving target detection is a very useful capability for a long range sensor. The combination of synthetic aperture radar and moving target detection has the potential to produce high resolution ground imagery with superimposed moving target information. Unfortunately, using conventional imaging data for detecting moving targets leads to ambiguities in the targets' positions and velocities. By using a non-uniform pulse repetition interval, the proposed ground imaging/moving-target detection radar overcomes this limitation and allows the azimuthal data to be focused at any velocity of interest, while collecting data at the same average rate as a conventional synthetic aperture radar. This approach permits the flexible use of a multimode radar, relaxes the specifications of data acquisition systems, affords a degree of protection against electronic countermeasures and retains a large unambiguous range swath, but with the added complexity of processing the non-uniform samples.

Having said that the Nyquist limit can be exceeded by irregular sampling, which seems like a very good prospect, we must be clear what class of signals can be unambiguously identified from irregular samples. There is a crucial distinction between spectral analysis and waveform reconstruction.

For signal reconstruction the Shannon theorem does not apply but variants of it do. First, simple interpolation (low-pass) schemes will only work if the signal is bandlimited to less than half the average sampling rate. Secondly, and more substantially, if the Fourier transform (or the spectrum) of the signal is confined[2] to a set I of measure B then reconstruction is possible providing the average sample rate exceeds B and providing I is known. Usually I is a finite union of subintervals[3]; the signal is called a *multiband signal*, I is the *spectrum support*, and B is the *bandwidth*

[1] This paragraph is taken from Legg's thesis [120].

[2] We are counting positive and negative frequencies when calculating I and B.

[3] So we do not need any sophisticated measure theory.

and also the *Nyquist-Landau rate.* This result was proven by Landau [115, 116] and has been rederived in an elementary way by the present author [141]. See §2.4.2, and footnote[4].

For spectral estimation, the spectrum of the underlying waveform is uniquely determined from the spectrum of (an infinite set of) samples, regardless of sampling rate, provided that an *alias-free* sampling scheme is used. See [189, 145] in the first instance, and §2.5.4. With regular sampling, the spectrum of the observations consists of periodic replications of the underlying spectrum; this gives rise to aliasing, i.e. the observed spectrum does not uniquely determine the underlying one. With random sampling, the spectrum of the observations consists of one faithful copy, and one smeared-out copy, of the underlying spectrum. So regular sampling is, by definition, not alias-free. As an example, Poisson sampling, in which the sampling instants constitute a Poisson process, is alias-free.

Let us now consider in general terms the question of how to analyse a multiband signal, in which the bands are at unknown frequencies. This is exactly the same problem as is encountered in the DASP system (q.v.), so we shall consider it from that point of view. Let us assume that technology at present imposes an upper limit of 1.2GHz, and that sampling is carried out at an average rate of 80MHz. Then from what we have already said, sampling theory and technology combined impose the following constraints on the system:

- It may unambiguously identify frequencies up to 1.2GHz, but no higher.
- It may reconstruct a signal bandlimited to 1.2GHz provided that the signal does not occupy the whole band 0–1.2GHz; specifically, the total width of its constituent bands[5] must not exceed 40MHz (half the Nyquist-Landau rate).

We are not saying that the present system can actually perform such tricks; these are simply the limits.

Assume that the input signal obeys the Nyquist-Landau constraint. Low-pass reconstruction methods, or simple interpolation, will fail because there are signal components at frequencies above half the average sampling rate (40MHz). Multiband reconstruction theory is useless until we have found I, the spectrum support. From what we have said about spectral estimation, though, it should be possible to obtain the spectrum from the samples (provided we used an alias-free sampling scheme). Then we would have I, and then we could apply the multiband reconstruction theorem. Unfortunately there is a problem with the spectrum estimation. Although alias-free sampling schemes permit unambiguous identification of frequency components, they cause the observed spectrum to be smeared so that only the strongest components are visible. Accordingly one cannot find I with any degree of certainty, except by making the unrealistic assumption that all the frequency components are

[4] In [141] the author suggests and shows that as the Nyquist-Landau rate is often a good deal less than twice the highest frequency present, irregular sampling offers great flexibility in sampling multiband signals from sensors. The same point is made by other authors [84].

[5] Here we are only talking about positive frequencies, because the signals are necessarily real-valued.

of roughly equal strength. (There is also the problem of *inconsistency*: the spectrum estimated from a finite data record (length N) does not settle down as $N \to \infty$.) It is therefore a question of sequentially identifying the frequency components in the spectrum and filtering them out of the time series. The difficulty there is that the theory of linear filtering is poorly developed if the data are irregularly spaced. Recall that we, and Summers [209], made this point earlier when asking why irregular sampling is not in greater use in sampled-data systems. Present techniques for this sequential-component extraction require the components to be sinusoidal so that they can be parametrised as $A\cos\omega t + B\sin\omega t$, and the parameters A, B found by linear least-squares [20, 69]. It follows that filtering is a key problem to be solved.

1.2 This thesis: its aims and themes

Aims

Combining questions of sampling theory with those of signal analysis gives the following list of questions. The aims of 'irregular sampling theory' are to answer them.

The aims of this thesis are to help answer these questions.

Q1 How do irregular samples of an underlying waveform relate to that waveform; for example do they, under appropriate sets of hypotheses, permit unique reconstruction?

Q2 Given an irregularly sampled signal, how should we process it?

Q3 How should we generalise common signal processing techniques such as filtering, convolution, model fitting, etc., to cope with irregularly sampled data?

The second two are closely related and cannot really be treated in isolation from each other. None of the questions is easy to answer.

Chapter 2 fills the need for a concise discussion of sampling theory and signal analysis that includes stochastic and nonlinear approaches, as well as the areas of complex analysis and integral transforms that are more familiar to 'pure' sampling theorists. It gives an overview of the (already large) literature on the subject, and finishes by drawing out some specific technical questions.

Chapter 3 is concerned with Q1, Q2 and Q3. §3.1 looks at specific types of sampling scheme, namely the sampling series associated with certain integral transforms such as the Hankel transform. It shows that filtering, convolution and system identification may be carried out using these irregular sampling schemes, provided that the theory is developed in a consistent way. Next, as there are explicit sampling series associated with these integral transforms it is desirable to have simple and tight truncation error bounds for them so that their practical value can be estimated. §3.2 examines this problem for the J_0-Bessel sampling series. Finally §3.3 has a different flavour. It is also focused on a specific type of sampling scheme—regular sampling with missing data—and shows that, assuming a simple condition on the sampling

set $\mathbb{L} \subset \mathbb{Z}$, a signal generated by harmonics[6] only has to be sampled at points on $\mathbb{L}$ to determine its values at the other integer points. This tells one how an infinite sensor array may be thinned with no (theoretical) loss in performance. The proof requires sampling theory for local fields.

Chapter 4 helps to answer Q2 and Q3; it rests on the idea of an underlying continuous-time harmonic or stochastic process, and so also addresses Q1. It looks in detail at the notion of 'generalised linear prediction' of linear signal analysis, which rather than attempting to express an observation as a linear combination of p previous values (with time-varying weights), seeks instead to treat $p+1$ consecutive values as linearly dependent. Extensive tests have shown that this is a robust technique and test results for signals with line spectra and with continuous spectra are shown. A natural progression of this work leads to the discussion of fast optimal filtering algorithms, thereby going a long way towards solving what we regard as the most important part of Q3.

Chapter 5, which is also concerned with Q2 and Q3, shows that signals that arise from dynamical systems may be processed using nonlinear prediction techniques even when the sampling is irregular. The underlying idea is to recognise determinism in a time series and use it to assist in further processing. It is shown that this substantially improves on linear methods for chaotic signals such as the Lorenz attractor. We also suggest how nonlinear filters should be constructed.

Appendices cover the derivation of the spectrum of an irregularly sampled signal (this has been done before but several of the oft-quoted texts [189, 143] contain crucial errors), and an introduction to some of the ultrametric techniques needed in the last part of Chapter 3.

The work in Chapter 4 is to be published in two journal papers [133, 136]. Other papers on the work in Chapters 3,5 are in progress.

Themes

The principal theme in this thesis is the analysis and processing of irregularly sampled data without recourse to explicit signal reconstruction. There are two facets to this idea, which we now discuss.

- We have already said that in the case of multiband signals with unknown frequency bands, one must not start one's analysis with an attempt at reconstruction. This was because we said that reconstruction cannot be accomplished correctly if there is no knowledge of the spectrum. The first step must be to find the bands. But there is another objection to immediate reconstruction: we may not want the waveform. For example we may only wish to answer a question similar to this, which occurs in radar signal processing:

  > Is this signal just white noise, or is it white noise plus a sinusoid?

  One can answer that question just by running a DFT on the samples. Reconstruction has no part to play, and it will make matters much worse if a bad reconstruction (based on erroneous spectrum information) is attempted.

[6] Actually the theorem is rather more general than that.

- Suppose that we have a signal processing technique that is defined for continuous-time signals. This is a very general concept; here is a specific example, the 'power function':

$$\Theta_r^c : x \mapsto \int_{-\infty}^{\infty} x(t)^r \, dt \qquad (r \in \mathbb{N})$$

Suppose that a sampled signal is given to us, that we know how to reconstruct the underlying signal from its samples via the function ρ say, and that we wish to calculate $\Theta_r^c(x)$ from the samples. For example, if the underlying signal is bandlimited to $\frac{1}{2}$ (frequency units) then ρ is given by (see §2.1.1)

$$\rho : (x_n)_{n\in\mathbb{Z}} \mapsto \left(t \mapsto \sum_{n\in\mathbb{Z}} x_n \operatorname{sinc}(t-n) \right).$$

It seems labouring an obvious point to say that we simply have to apply ρ and then Θ_r^c :

$$\Theta_r^d (= \Theta_r^c \rho) : (x_n)_{n\in\mathbb{Z}} \mapsto \int_{-\infty}^{\infty} \left(\sum_{n\in\mathbb{Z}} x_n \operatorname{sinc}(t-n) \right)^r dt. \tag{1.1}$$

But there is a tendency[7] not to do this, preferring instead the simpler

$$\Theta_r^d : (x_n)_{n\in\mathbb{Z}} \overset{??}{\mapsto} \int_{-\infty}^{\infty} \left(\sum_{n\in\mathbb{Z}} x_n \delta(t-n) \right)^r dt.$$

The realisation dawns that $\int \delta(t-n)^r \, dt$ is meaningless when $r \geq 2$. In fact $g(\delta(t))$ never is well-defined except in trivial cases (because it is not a distribution). But pressing on undeterred, the practitioners write $g(\delta(t)) \equiv g(1)\delta(t)$ and are now able to conclude (?)

$$\Theta_r^d : (x_n)_{n\in\mathbb{Z}} \overset{??}{\mapsto} \sum_{n\in\mathbb{Z}} x(n)^r. \tag{1.2}$$

This is consistent with (1.1) when $r = 1$, and also when $r = 2$ remarkably enough (because the sinc functions are orthonormal on $\mathbb{R}$); but not[8] when $r \geq 3$.

The reason why the delta-function approach fails is that it ignores the whole basis of sampling and reconstruction—namely that a continuous-time signal is uniquely determined by its samples, when certain assumptions hold, via a reconstruction formula.

This argument might at first appear contradictory to the theme of analysing and processing data without recourse to explicit reconstruction; but it is not. The key word is *explicit.* The expression (1.1) is an example of reconstruction being used *implicitly* and *consistently.* If we write down Θ_r^d as a function of the samples, and tidy up the algebra, we have avoided explicit reconstruction, and we have a formula consistent with the ideas we had when we did the sampling. That is what §3.1 is about.

[7] Private communications to the present author.

[8] Let $x : t \mapsto \operatorname{sinc} t$. Then $\Theta_r^c(x) \to 0$ as $r \to \infty$. But $x_n = 1$ for $n = 0$, and $x_n = 0$ otherwise. So the RHS of eq.(1.2) is 1 for all r and so doesn't tend to 0 as $r \to \infty$.

Relation of Chapter 2 to existing texts

Several books on sampling theory have emerged in the past few years and all have different emphases. Marvasti's book [143] undertakes an in-depth review of the theory of zero-crossings and related issues. He also provides an extensive overview of the literature (particularly applications of nonuniform sampling). Bilinskis' account [23] is rather different, concentrating on the advantages that deliberate irregular or random sampling may have[9] with regard to aliasing and his is the only text to do so extensively; his approach is almost to regard uniform sampling as a rather inconvenient special case. Marks II [127] has compiled a large text in which leading authorities have written chapters on their various subjects, such as Gabor analysis, optics, multidimensional sampling, sampling in polar coordinates, and error analysis; its bibliography is substantial, running to over 1000 references. Zayed's account [229] is technical, concentrating on the Shannon and Kramer-Weiss sampling theorems. Finally Higgins' superb text [84] is easily the best if one seeks a comprehensive discussion of sampling theory written from a mathematical standpoint, combining insight with mathematical precision to produce a balanced and very readable account of the subject.

1.3 Notation

As this thesis uses ideas and results from different branches of mathematics and engineering the notation has been standardised; that adopted here necessarily sometimes differs from that to which specialists are accustomed. The following is not an exhaustive list of notation used but is intended to clear up any confusion before it should arise.

Chapter 2

The Fourier transform is always written in the 'engineering notation'

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} x(t)e^{-2\pi ift}dt \\ x(t) &= \int_{-\infty}^{\infty} X(f)e^{2\pi ift}df \end{aligned}$$

as opposed to $\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt$ which is usual in the sampling theory literature. The chosen form is symmetric and more useful for making remarks about time-frequency reciprocity.

A signal is *bandlimited* to W iff its Fourier transform vanishes for $|f| > W$.

A signal has *bandwidth* B if its Fourier transform is supported on the set I of measure B. Hence an arbitrary signal bandlimited to W has bandwidth $2W$, not W. The reader should be careful about this.

[9]—both in the time domain (abscissa) and the 'amplitude domain' (ordinate), the latter being the subject of irregular quantisation levels. Some mention of Monte Carlo techniques is made, but Hammersley & Handscomb's fine text [78] is missing from the Bibliography.

Chapter 3, Section 1

This discusses the so-called K-transform, which would be better described as an integral representation of a signal. It is

$$x(t) = \int X(u)K(u,t)\rho(u)\,du.$$

K denotes the kernel; $f(t)$, $x(t)$ etc. denote functions of time with transforms $F(u)$, $X(u)$ etc.; $\ominus$ and $\circledast$ are the generalised shift and convolution operators.

Chapter 3, Section 2

τ denotes complex time and its real and imaginary parts are u, v. Hence $x = x(\tau) = x(u + \mathrm{i}v)$ rather than the more normal $f = f(z) = f(x + \mathrm{i}y)$.

Chapter 3, Section 3

$\mathbf{Z}_p$ is the ring of p-adic integers (*not* the rational integers taken modulo p, as it sometimes means): see Appendix B.

Chapter 4

$\mathcal{E}^x$ denotes the expectation taken over all realisations of the random process $x(t)$ (realisation average). $\mathcal{E}_n$ denotes the expectation taken over time n (time average). $B(t)$ is a Brownian motion process, i.e. random walk obeying the condition

$$\mathcal{E}^B dB(t_1)^* dB(t_2) = \begin{cases} \sigma_B^2 dt, & t_1 = t_2 \\ 0, & t_1 \neq t_2 \end{cases}$$

where $dB(t_i)$ are increments of the process at times t_i over an infinitesimal time dt. The quantity σ_B^2 is known as the diffusion coefficient and has dimensions $[B]^2[T]^{-1}$ where B, T denote the dimensions of the physical variable and of time.
τ is the effective intersample spacing (or, equivalently, $1/2\tau$ is the 'normalising frequency' playing the role of the classical Nyquist frequency).
z denotes the transform-variable in the z-transform; $\mathbf{z} = \{z_n\}$ is a time series.

Chapter 5

τ_n denotes the intersample spacing $t_{n+1} - t_n$.
Underlining denotes a 'delay-vector', i.e. a vector of d consecutive observations viewed in $\mathbf{R}^d$.

Chapter 2

Overview

2.1 The Shannon (WKS) sampling theorem

2.1.1 Introductory remarks

The sampling theorem referred to here was introduced by Shannon to information theory [188]. However, the interest of the communications engineer in sampling and reconstruction can be traced back to Nyquist [156]. The theorem was originated by E.T. and J.M. Whittaker, W.L. Ferrar and V.A. Kotelnikov. In an extensive review of various aspects of the Shannon sampling theorem and its extensions, A.J. Jerri [94] refers to it as 'WKS', after these authors. Other reviews have been carried out by Butzer [30, 31, 32, 34, 38] and Higgins [82, 84].

Shannon's original statement [188] of the sampling theorem was:

> If a function $x(t)$ contains no frequencies higher than W it is completely determined by giving its ordinates at points spaced $1/2W$ apart.

Its significance was that for the first time one could see that a bandlimited signal is uniquely determined by its samples via an explicit formula (which we shall now derive) and for an explicit sampling rate. Shannon's proof starts by letting

$$x(t) = \int_{-\infty}^{\infty} X(f)e^{2\pi ift}df = \int_{-W}^{W} X(f)e^{2\pi ift}df, \tag{2.1}$$

the second equality holding because $X(f)$, the Fourier transform of $x(t)$, defined as

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-2\pi ift}dt,$$

is assumed to be zero outside the range $[-W, W]$. Next $X(f)$ is expanded as a Fourier series on the interval $(-W, W)$ as

$$X(f) = \sum_{n\in\mathbb{Z}} c_n e^{-2\pi ifn/2W} \tag{2.2}$$

and the coefficents c_n obtained by Fourier analysis:

$$c_n = \frac{1}{2W}\int_{-W}^{W} X(f)e^{2\pi i f n/2W}df = \frac{1}{2W}x\left(\frac{n}{2W}\right) \text{ by (2.1)} \tag{2.3}$$

We note that c_n is essentially $x(n/2W)$, the sample of the signal x at the point $t = n/2W$. Finally we substitute (2.3) into (2.2) into (2.1) and perform the integration to get

$$x(t) = \sum_{n\in\mathbb{Z}} x\left(\frac{n}{2W}\right)\text{sinc}\,(2Wt-n) = \sum_{n\in\mathbb{Z}} x\left(\frac{n\pi}{a}\right)\frac{\sin(at-n\pi)}{at-n\pi} \tag{2.4}$$

with

$$a = 2\pi W$$

(a notation which will be used from now on).

One sees from the above proof of the WKS theorem that the functions

$$B_n(t) = \text{sinc}\,(2Wt-n)$$

form an orthogonal basis for the space of signals bandlimited to W : the spanning property is provided by (2.4) and the orthogonality follows from that of the Fourier basis employed in (2.2), by Plancherel's identity. Note also that $B_n(t)$ is 1 at the nth sample point and 0 at the others, so *all the samples are necessary.* Given that the independent samples arrive every $1/2W$ and that the bandwidth is $2W$, we arrive at the conclusion that Bandwidth = Information rate (in some sense). This is the basis of Shannon's '$2WT$' theorem which states that in a channel bandlimited to W one can only transmit $2WT$ independent numbers in a time T : see e.g. [188, 193].

Multidimensional analogues, which are of interest because of the different sorts of sampling geometry and lattices that one can obtain, are given in [49, 197, 84].

2.1.2 Aliasing

For the WKS theorem to hold we require $X(f)$ to vanish outside[1] $[-W,W]$. If it does not, the reconstruction formula can be applied but an error occurs, known as the aliasing error:

$$\epsilon_A(t) = x(t) - \sum_{n\in\mathbb{Z}} x\left(\frac{n\pi}{a}\right)\frac{\sin(at-n\pi)}{at-n\pi}.$$

System-theoretically the aliasing error can be understood as follows. When a signal is sampled at rate R, the transform of the sampled signal consists of a sum of displaced copies of the transform of the continuous-time waveform; the displacements in f-space are multiples of R. If $X(f)$ does not vanish outside the range $[-\frac{1}{2}R,\frac{1}{2}R]$, the copies overlap and the reconstruction process is not able to remove the displaced copies. It is apparent therefore that only the out-of-band spectral density should

[1] We assume X to be continuous at $\pm W$. This is to prevent X having a delta-function contribution at the ends. For example if $x(t) = \sin \pi t$ and $W = \frac{1}{2}$, then the samples are all 0 and the reconstruction fails.

contribute to $\epsilon_A(t)$, and indeed the following bound, derived by Stickler [206] in a less transparent form, has this property[2]:

$$|\epsilon_A(t)| \leq 2 \sum_{k \in \mathbb{Z}} |\sin kat| \int_{-W}^{W} |X(f + 2kW)|\, df.$$

For the proof, take the previous equation, Fourier transform everything in sight, and use the Poisson summation formula, $\sum_n e^{2\pi i n u} \equiv \sum_m \delta(u - m)$. (For a discussion of the PSF, see e.g. [84, 121, 98, 13], and a nice paper by Schmeisser [182].) This estimate has the merits of vanishing at the sample points (as it should) and being tight in several nontrivial cases, e.g. when $x(t) = e^{2iat}$. Splettstösser [196] gives a useful overview of aliasing error bounds.

2.1.3 Multiband signals

The discussion on aliasing and derivation of the reconstruction kernel $\mathrm{sinc}\,(t)$ suggest a substantial generalisation of the WKS theorem, as follows. Suppose that $X(f)$ is supported on (i.e. vanishes outside) a set I of measure $B < \infty$ (technically this means the Lebesgue measure, but we shall only be dealing with very simple-minded sets I) and that sampling is carried out at rate R. First, if translates of I through all integer multiples of R are disjoint, then there is no obvious source of aliasing. Secondly, we can write (2.4) as

$$x(t) = \sum_{n \in \mathbb{Z}} x(n/R)\, \mathrm{ke}\,(t - n/R) \tag{2.5}$$

with $\mathrm{ke}\,(t)$ the inverse Fourier transform of the indicator function of I, up to a factor:

$$\mathrm{ke}\,(t) = R^{-1} \int_{-\infty}^{\infty} 1_{f \in I} e^{2\pi i f t} df.$$

This formula is meaningful whether or not I is in one piece so does it work when the condition of disjointness holds? Remarkably the answer is 'yes'; loosely, the proof consists in shifting the various parts of I through multiples of R, until they lie in the interval $[-R, R]$, and then applying (2.4); see Higgins [84] §13.

Having established that (2.5) holds we must be clear what the admissible values of R are, given I. A little thought shows that R must exceed B, the so-called *Nyquist-Landau rate*. (It turns out that even if the sampling is irregular the minimum average sampling rate must exceed B; see §2.4.) An algorithm for finding the minimum admissible value of R is given by Dodson & Silva [55].

A simple case that admits direct analysis is the 'bandpass sampling' of a signal bandlimited to $W_l \leq |f| \leq W_u$. In that case the admissible values of R are those for which no integer n satisfies the inequality $nR \in (2W_l, 2W_u)$ —a well-established result [172]. Furthermore the lowest such R (R_0 say) is $2W_u/\lfloor W_u/(W_u - W_l) \rfloor$, with $\lfloor \cdot \rfloor$ denoting the integer part; in particular if $W_l/(W_u - W_l) \in \mathbb{Z}$ then R_0 is the Nyquist-Landau rate $2(W_u - W_l)$. Provided that R is admissible, we have (2.5) with

$$\mathrm{ke}\,(t) = \frac{\sin 2\pi W_u t - \sin 2\pi W_l t}{\pi R t}.$$

[2] And implies the older and better-known result $|\epsilon_A| \leq 2 \int_{|f|>W} |X(f)|\, df$.

When the sampling is random the reconstruction process is more difficult, but the choice of sampling rate is more flexible, as we can choose any $R > B$.

2.1.4 Self-truncating sampling expansions

The concept of a self-truncating sampling expansion was introduced by Helms & Thomas [81] with the view to increasing the rate of convergence of the sampling series. Suppose that $x(t)$ is bandlimited to rW, for some $r < 1$, but that samples are still taken at rate $2W$. The kernel[3]

$$h_m(t) = \left[\operatorname{sinc} \frac{2qWt}{m}\right]^m$$

is decaying, bandlimited to qW, and is 1 at $t = 0$. Pick $q = 1 - r$ and apply the WKS theorem to $x(t)h_m(t' - t)$, which has bandlimit $rW + qW = W$:

$$x(t)h_m(t' - t) = \sum_n x\left(\frac{n}{2W}\right)\left[\operatorname{sinc} \frac{2qW}{m}\left(t' - \frac{n}{2W}\right)\right]^m \operatorname{sinc}(2Wt - n)$$

and put $t' = t$:

$$x(t) = \sum_n x\left(\frac{n}{2W}\right)\left[\operatorname{sinc} \frac{q}{m}(2Wt - n)\right]^m \operatorname{sinc}(2Wt - n). \tag{2.6}$$

This gives a more rapidly convergent series than (2.4). A system-theoretic approach shows why. The reconstruction process corresponds to multiplication by a gate-function ($1/2W$ if $|f| \leq W$, 0 if not) in the frequency domain, and the discontinuity in this function causes the impulse response to decay slowly (as t^{-1}). If the signal is known to have an empty spectrum between $(1 - q)W$ and W, the spectrum of the samples will be empty between $(1-q)W$ and $(1+q)W$, and the filter transfer function can be arbitrary in that band: in particular it can be made so that it drops smoothly from 1 (at frequency $(1 - q)W$) to 0 (at $(1 + q)W$). Here the transfer function is a convolution of gate-functions. A related point is of interest: the reconstruction kernel is bandlimited to $(1 + q)W$ rather than W.

Helms & Thomas then considered the possibilities of adjusting m to minimise the truncation error, i.e. the error induced by neglecting the terms with $|n| >$ some N. In the same paper they discussed a contour integral approach for the estimation of truncation errors (this will be discussed in §2.3) and after lengthy calculations gave an approximate 'optimal' value of m. We shall show here that judicious guesswork gives the same answer. Near the origin the excluded terms in (2.6) are approximately

$$\sum_{|n|>N} x\left(\frac{n}{2W}\right)\left[\operatorname{sinc} \frac{qn}{m}\right]^m \frac{\sin 2\pi Wt}{\pi n}.$$

The dominant contribution comes from the lowest value of $|n|$ in the summation, and is of order $(m/Nq\pi)^m$ (as $\operatorname{sinc} u$ decays as $1/\pi u$). This expression assumes its minimum at (the integer closest to)

$$m_0 = Nq\pi/e$$

[3]The same applies if in h_m we replace $\operatorname{sinc}(\cdot)$ by $J_1(\pi\cdot)/(\pi\cdot)$, where J_1 is the Bessel function.

and this is the same result as given in [81]. The truncation error is reduced by a factor of $\approx (e^{\pi/e})^{Nq}$, which is quite substantial (e.g. N=10, $q = 0.5$ cause a 300-fold reduction).

2.1.5 Zhu's theorem

A so-called generalisation of the WKS theorem has been enunciated by Zhu [230], namely that even if a signal $y = y(t)$ is not bandlimited to W, it may be recovered by sampling at rate $2W$ if there exists an invertible transformation $g : \mathbf{R} \to \mathbf{R}$ such that $gy = g(y(t))$ *is* bandlimited to W. One simply reconstructs gy using the sampling theorem and then applies g^{-1}. The intention is that g is a smooth invertible function, which will have to be nonlinear if it is to do something interesting.

Now a nonlinearity nearly always increases the bandwidth of a signal, and one is led to ask for an example of a bandlimited signal y and invertible g such that gy has a lower bandlimit than y. An example does arise if $y = hx$ for some bandlimited x and smooth invertible h. At that point one realises that Zhu's theorem is stated 'inside-out'. The practical version runs as follows. Let $x = x(t)$ be bandlimited to W. Then it can be recovered from instantaneously-distorted samples $h(x(n/2W))$ by applying h^{-1} and then (2.4).

Like Zhu's result this is also trivial but it has an important practical consequence, which is that an instantaneous nonlinearity may be identified and corrected at the Nyquist rate of the *input*. Recent work [70] has shown that nonlinearities with memory—cast in the form of Volterra operators—may also be identified when the input and output are both sampled at the input Nyquist rate. We shall discuss this, from the point of view of the Kramer-Weiss sampling theorem, in Chapter 3.

2.2 Kramer's sampling theorem

2.2.1 Introductory remarks

Looking at the proof of the (WKS) sampling theorem we see that it requires the following ingredients:

- A signal that is transform-limited, $x(t) = \int_I X(u)K(u,t)\rho(u)\,du$ for some kernel K and nonnegative weight function ρ
- A countable subset $\{\ell_n\}$ of $\mathbf{R}$, indexed by $\mathbf{X}$ say, such that $K(u,\ell_n)$ are orthogonal functions of u on I,

$$\int_I K(u,\ell_m)^* K(u,\ell_n)\rho(u)\,du = 0 \quad (m \neq n)$$

  Technically we require $\{K(\ell_n) : n \in \mathbf{X}\}$ to be an orthogonal basis for $L^2(I)$.
- An expansion, $X(u) =$ linear combination of the $K(u,\ell_n)$'s.

Notice that, despite the term 'integral transform', $X(u)$ is not obtained from $x(t)$ by an integral; it is the other way round, and maybe 'integral representation' would

be a more appropriate term. Kramer's sampling theorem [113, 222] asserts that if $X(u)$ is zero outside I then $x(t)$ can be reconstructed from its samples $x[n] = x(\ell_n)$ by

$$x(t) = \sum_{n \in \mathbb{X}} x[n] \frac{\int_I K(u,t) K(u,\ell_n)^* \rho(u)\, du}{\int_I |K(u,\ell_n)|^2 \rho(u)\, du}. \tag{2.7}$$

The proof is identical to that of the WKS theorem (§2.1.1). Indeed to obtain the WKS theorem from (2.7) we simply put

$$K(f,t) = e^{2\pi i f t}, \quad \rho(f) \equiv 1/2W, \quad I = [-W, W], \quad \ell_n = \frac{n}{2W}.$$

Other transforms that are encompassed by this definition are the Hankel (Bessel), Legendre and Chebyshev transforms [228, 7, 42, 40, 152, 99, 93, 94]. For example, in the case of the J_0-Hankel transform,

$$K(u,t) = J_0(ut), \quad \rho(u) = u, \quad I = [0, b], \quad \ell_n = j_{0,n}/b$$

($j_{0,n}$ denotes the nth zero of $J_0(t)$) and the expansion is

$$x(t) = \sum_{n=1}^{\infty} x\left(\frac{j_{0,n}}{b}\right) \frac{2 j_{0,n} J_0(bt)}{(j_{0,n}^2 - b^2 t^2) J_1(j_{0,n})}$$

in which we have used common properties of the Bessel functions [2, 221].

Initially the kernels were taken from self-adjoint (Sturm-Liouville) boundary value problems, which are mathematically 'pleasant' to deal with (eigenvalues real, eigenfunctions orthogonal, etc.). For example, the Fourier transform comes from the system with operator

$$L(v) = -\mathrm{i}\frac{dv}{dt}$$

and the J_m-Hankel transform from the system with (singular) operator

$$L(v) = -\frac{d^2 v}{dt^2} + \frac{m^2 - \frac{1}{4}}{t^2} v.$$

See Zayed's and Jerri's texts [229, 97] for further details of sampling theorems from differential operators, and also [85]. Recent work by Annaby [4, 5] concerns more general sources of transform kernels and their 'resolvent' kernels. The question of what operators are 'permitted' remains open. Other work [6, 71, 72] considers sampling theorems associated with discrete transforms and difference equations.

2.2.2 Further developments

We note first that when X vanishes outside I, it can be obtained directly from the time-domain samples $x[n] = x(\ell_n)$, by the discrete-time K-transform or 'DKT':

$$X(u) = \sum_{n \in \mathbb{X}} x[n] \frac{K(u,\ell_n)^*}{\kappa[n]} \tag{2.8}$$

where

$$\kappa[n] = \int_I |K(u, \ell_n)|^2 \rho(u)\, du.$$

(Jerri [95, 98] derives this for the case of the Hankel transform.) For the proof, simply compare

$$\int_I X(u)K(u,t)\rho(u)\, du = \sum_{n\in\mathbf{X}} \frac{x[n]}{\kappa[n]} \int_I K(u,t)K(u,\ell_n)^*\rho(u)\, du.$$

From (2.8) we have a discrete version of the Plancherel formula:

$$\int_I X(u)^* Y(u)\rho(u)\, du = \sum_{n\in\mathbf{X}} \frac{x[n]^* y[n]}{\kappa[n]} \tag{2.9}$$

or $\langle X, Y\rangle = \langle x, y\rangle$ for short. Consider now the interpretation of a convolution for the K-transform. We are used to the standard Fourier convolution being a multiplication in the transform domain, so we define a K-convolution by [98, 88, 54]

$$(f \circledast g)(t) \equiv \int F(u)G(u)K(u,t)\rho(u)\, du$$

It is also possible to define a K-translation [98]:

$$f(t \ominus \tau) \equiv \int F(u)K(u,t)K(u,\tau)^*\rho(u)\, du.$$

Note that on its own $t \ominus \tau$ does not mean anything; $\ominus$ actually operates on the function f. Some caution is required in the manipulation, for $f(t \ominus t)$ does not mean the same as $f(0)$; also the K-translations do not in general form a group. The Fourier case is quite familiar, for then the K's are just exponentials (and they do form a group), $\ominus$ gives rise to a straightforward translation, and the above equation simply says that a translation in one domain corresponds to a modulation in the other. Using the generalised translation we can obtain a rather neat statement of Kramer's sampling theorem:

$$x(t) = \sum_{n\in\mathbf{X}} \frac{x[n]}{\kappa[n]} \mathrm{ke}\,(t \ominus \ell_n)$$

in which the reconstruction kernel ke has K-transform $1_{u\in I}$.

A final point concerns the DKT. For an arbitrary signal $x(t)$, whose transform is $X(u)$, we have

$$x(t) = \int X(u)K(u,t)\rho(u)\, du.$$

Consider now

$$X^s(u) = \sum_{n\in\mathbf{X}} x[n] \frac{K(u,\ell_n)^*}{\kappa[n]}.$$

Suppose that x is transform-limited. Then X^s and X are identical on I, but not outside (as X vanishes, but X^s doesn't). In general $X^s(u)$ can be thought of as a convolution of $X(u)$ with a *generalised impulse train* (array of delta-functions at certain positions and of certain strengths, which depend on the transform kernel K). Equivalently, X^s consists of generalised translations of X (in the Fourier case these would be straightforward translations). Using this device Jerri obtains a bound for the aliasing error of the J_0-Hankel transform [96, 98].

2.3 Truncation errors and contour integration

All sampling expansions involve infinitely many terms, so one often wants to estimate the truncation error, i.e. the error obtained by taking only finitely many terms, as is always the case in practice. The earliest workable bound for the WKS expansion was given by Helms & Thomas [81, 224], which apart from being in closed form has the advantage of offering considerable insight into sampling and reconstruction problems and providing a general technique by which more general sampling expansions may be derived. See [94] for a review of truncation error bounds, and [35, 83] for discussion of the role of complex analysis in sampling theory.

We first note that a bandlimited signal is entire (analytic and free from singularity), because it can be written

$$x(\tau) = \int_{-W}^{W} X(f) e^{2\pi i f \tau} df \tag{2.10}$$

which is valid for all $\tau \in \mathbb{C}$. This incidentally leads to some quite interesting observations about bandlimited functions: for one, using the Identity Principle for analytic functions, we see that no bandlimited function can be timelimited too (unless it vanishes identically), and that knowledge of a bandlimited function in some interval is sufficient to define it all over $\mathbb{C}$. As Marks points out ([126], p.257) this second point can if misinterpreted lead to some alarming conclusions: for example, a telephone conversation can be considered bandlimited, and so can be determined if we know only a word or two in the middle (!). See [84]§17, [193], and footnote[4].

Returning to (2.10) we see that x is of exponential-type $a = 2\pi W$, in that $|x(\tau)| < Ke^{a|\tau|}$ on $\mathbb{C}$. Remarkably the converse is also true, in that an entire function of exponential type (EFET) must necessarily be bandlimited (the Paley-Wiener theorem [159]). We may therefore work with EFETs and use methods of complex variable theory; τ may be thought of as 'complex time'. Its real and imaginary parts will be called u, v.

We sketch the method of Helms & Thomas. The signal $x(t)$ is assumed to be bounded by M on the real axis, and to be of exponential type, but bandlimited to $[0, rW]$ rather than $[0, W]$, i.e. the signal is oversampled. By a theorem of Duffin & Schaeffer [57], we have $|x(u + \mathrm{i}v)| \leq M \cosh rav$ (if x is real on the real axis), or $|x(u + \mathrm{i}v)| \leq Me^{ra|v|}$ (if not). We shall use the former, as it makes the integrals easier in the ensuing discussion. Consider the integral

[4] Slepian [193] resolves this 'paradox' by providing a philosophical distinction between the underlying signal, which can never be perfectly observed, and one's model for it; concepts of bandlimitedness only apply to the model, so the question 'Are real signals bandlimited?' is not meaningful. This seems rather dubious. Consider the question of how 'unbandlimited' a time-limited signal has to be. The answer is 'Not very much at all'. Indeed if we consider the operators $\mathcal{B}$ and $\mathcal{T}$ that bandlimit and timelimit a signal to $|f| < \frac{1}{2}B$ and $|t| < \frac{1}{2}T$, we are looking at how close the eigenvalues of the operator $\mathcal{BT}$ can come to 1. The eigenfunctions of $\mathcal{BT}$ are the prolate spheroidal wavefunctions, discussed in a classic series of papers [162, 118, 119, 192, 194], and if the eigenvalues are $\lambda_0 > \lambda_1 > \cdots$ then $1 - \lambda_i$ scales exponentially with $-BT$. Slepian states that as copper wires cannot transmit frequencies above (say) 10^{20}Hz, a 'paradox' must occur. To knock over the paradox only requires us to allow something in the region of $10^{-10^{20}}$ of the total signal energy in frequencies $> 10^{20}$Hz. Can this really be objected to on physical grounds?

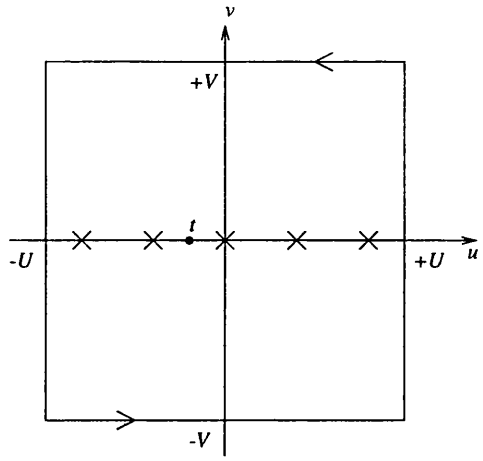


Figure 2.1. Integration contour for derivation of sampling series.

$$I_{\square} = \frac{1}{2\pi \mathrm{i}} \oint \frac{x(\tau)}{(\tau - t)} \frac{\sin at}{\sin a\tau} d\tau \tag{2.11}$$

around the box contour shown in Figure 2.1. The crosses are at zeros of $\sin a\tau$, and the samples at $(t = n\pi/a,\ -N \leq n \leq N)$ are inside the contour.
The sides $\pm U$ are chosen at zeros of the derivative of $\sin a\tau$. We let $V \to \infty$ and can now evaluate the integral in two different ways. First we consider the residue theorem, and find that

$$I_{\square} = x(t) + \sum_{|n| \leq N} x\left(\frac{n\pi}{a}\right) \frac{(-)^n \sin at}{n\pi - at} = x(t) - \sum_{|n| \leq N} x\left(\frac{n\pi}{a}\right) \frac{\sin(at - n\pi)}{at - n\pi}$$

which we identify as the truncation error: $I_{\square} = \epsilon_T(t)$. Secondly we consider the contributions from each side. On the top and bottom, we see that the numerator scales as e^{raV} and the denominator as e^{aV}, so these sections do not contribute. On the left and right sides, $|\sin a\tau| = \cosh av$ and so the contributions from these sections are bounded by

$$\frac{M|\sin at|}{2\pi|t \pm U|} \int_{-\infty}^{\infty} \frac{\cosh rav}{\cosh av} dv = \frac{M|\sin at|}{2a\cos(r\pi/2)|t \pm U|}$$

(By choosing U to be a zero of $\cos a\tau$, we minimise the contributions from the sides.) Writing $U = (N + \frac{1}{2})\pi/a$,

$$|\epsilon_T(t)| \leq \frac{M|\sin at|}{2\cos(r\pi/2)} \left\{ \frac{1}{|at - (N + \frac{1}{2})\pi|} + \frac{1}{|at + (N + \frac{1}{2})\pi|} \right\} \tag{2.12}$$

We make the following observations:

(a) Convergence. The truncation error bound tends to 0 as $N \to \infty$.

(b) Assumptions. $x(t)$ needs only to be an EFET and in $L^{\infty}(\mathbb{R})$. To make the proof of §2.1.1 rigorous requires $x(t)$ to be in L^2 (and then the sampling series is absolutely convergent, not just convergent).

(c) Bandwidths. We require $r < 1$; indeed the integrals will not converge if $r \geq 1$. This gives a succinct justification of the WKS theorem, as follows. The bandlimit on x gives the rate of exponential growth of $x(\tau)$ as $2\pi W$; the rate of sampling R gives the rate of exponential growth of the denominator as πR; for a valid expansion the latter must exceed the former.

(d) Generalisation. One can generalise to arbitrary sampling schemes, by replacing $\sin a\tau$ with another function ($S(\tau)$ say), the zeros of which correspond to the sample points. The residue theorem gives the *Lagrange interpolation formula*

$$x(t) = \sum_n x(t_n) \frac{S(t)}{(t - t_n)S'(t_n)} \tag{2.13}$$

Further, we may include sampling of the derivatives of x. If S has a zero of order $d+1$, the sampling expansion will contain the value of $x(t)$ at that point and the values of its $1, 2, \ldots, d$th derivatives. Indeed the exponential-type of $S(\tau)$ is related closely to the density of its zeros, and there is a large body of work devoted to this matter, including several books (see [171] in the first instance). The question of what the 'allowable' sampling schemes are, i.e., those that make (2.13) correct, is a delicate one discussed later in §2.4.2.

(e) Extrapolation. The truncation error bound increases to ∞ at the ends of the observation interval, suggesting that the formulae given here cannot be used for extrapolation. Now a bandlimited function *can* be extrapolated from past samples (though the problem is very unstable) provided that it has been 'oversampled', i.e. sampled at a rate above the minimum rate prescribed by the sampling theorem. The reason why the above contour integral does not do the job is that, despite the assumption that x is bandlimited to rW, the sampling series uses reconstruction functions of the form $\operatorname{sinc}(2Wt - n)$, which are linearly independent when the sampling rate is $2W$. The extrapolation requires an expansion using the functions $\operatorname{sinc}(2rWt - n)$, which are overcomplete (dependent) at sample rate $2W$.

(f) Two-sidedness. For arbitrarily good accuracy of reconstruction at a point t, one needs to take infinitely many samples on *both* sides of t. There has been interest in reconstruction from past samples [39], in which one needs only to take samples on one side of the point in question. We show in [140] that this may be achieved by the contour integral method, using a kernel of the form $\Gamma(z)e^{-\lambda z}$.

2.4 Reconstruction

There is a large body of work devoted to reconstructing signals from nonuniform samples. In the first two subsections we deal with methods that, despite the use of a finite number of samples, do produce a bandlimited reconstruction. Other methods, such as sample-and-hold techniques, splines, and finite-order Lagrange interpolation, are simplifications of the fundamental sampling theorems, and are discussed in the third.

2.4.1 Using the WKS (sinc) kernel

Following on from remark (e) above, we now turn to more general problems of reconstruction, when the restriction of bandlimitedness is retained but the sample points assume arbitrary positions. If the bandlimit is W then the signal can be reconstructed as a linear combination of any of the functions $\sin a(t-\tau_m)/a(t-\tau_m)$ with $a = 2\pi W$ as before. The τ_m are arbitrary. Now it is plain that if there are only a finite number of observations to be interpolated then there are infinitely many reconstructions of the form

$$x^r(t) = \sum_m \xi_m \frac{\sin a(t-\tau_m)}{a(t-\tau_m)} \tag{2.14}$$

obeying the interpolation condition

$$x^r(t_n) = x_n. \tag{2.15}$$

It has been shown (in [226, 15]; see also [48, 41, 231, 211]) that the minimum-energy solution, i.e. that which minimises $\int |x^r(t)|^2 dt$, uses sinc functions 'centred' at the observation points, i.e.

$$\tau_n := t_n \qquad \text{(Minimum norm)}.$$

The result, usually attributed to Yen [226], may be obtained by minimising the signal energy subject to (2.15), but a rather more elegant approach may be taken, as follows. Consider two possible reconstructions, $x(t)$ and $x^\circ(t)$, defined by

$$x^\circ(t) = \sum_n \xi_n^\circ \frac{\sin a(t-t_n)}{a(t-t_n)}, \qquad x(t) = \sum_n \xi_n \frac{\sin a(t-\tau_n)}{a(t-\tau_n)}$$

and both interpolating the observation points, so $x^\circ(t_m) = x(t_m) = x_m$. Then

$$\begin{aligned} \|x - x^\circ\|^2 &= \|x\|^2 + \|x^\circ\|^2 - \frac{\pi}{a} 2\mathrm{Re} \sum_{m,n} \xi_n^\circ \xi_m^* \frac{\sin a(\tau_m - t_n)}{a(\tau_m - t_n)} \\ &= \|x\|^2 + \|x^\circ\|^2 - \frac{\pi}{a} 2\mathrm{Re} \sum_{m,n} \xi_n^\circ \xi_m^{\circ *} \frac{\sin a(t_m - t_n)}{a(t_m - t_n)} \\ &= \|x\|^2 + \|x^\circ\|^2 - 2\|x^\circ\|^2 \end{aligned}$$

Therefore

$$0 \le \|x - x^\circ\|^2 = \|x\|^2 - \|x^\circ\|^2$$

and x° is the minimum-energy reconstruction. In Chapter 3 we shall show that this method of proof generalises to the K-transforms. The above derivation requires the inner product of two of the basis functions:

$$\int_{-\infty}^{\infty} \frac{\sin a(t-t_1)}{a(t-t_1)} \frac{\sin a(t-t_2)}{a(t-t_2)}\, dt = \frac{\pi}{a} \frac{\sin a(t_1 - t_2)}{a(t_1 - t_2)}$$

This neat result is known as the Hardy integral and its generalisation to the K-transforms, as mentioned by Jerri [98], allows the minimum-norm proof to go through in the same way as above.

Elegant though Yen's result is, it cannot be regarded as a sensible way to reconstruct signals in practice. As we shall see later, low-pass reconstruction is stable only when the sampling rate is on average higher than twice the highest frequency present, thereby allowing overcompleteness in the set of 'basis functions' in (2.14). Given then that the 'basis functions' are no longer independent, we must expect the matrix inversion (needed to obtain the ξ_n's from the samples x_n in (2.14)) to be ill-conditioned (see [225]). One can use singular-value decomposition to identify and remove almost-zero combinations of generating functions [223]. A more satisfactory idea is to use a basis set that is irredundant to start with, and that is the key to methods currently used at the NUHAG group in Vienna University. By considering the periodic extension of the data and fitting a truncated Fourier series (i.e. using the transform kernel), the problem reduces to the inversion of a Toeplitz matrix, which can be tackled by a number of techniques to reduce the computational load. This can also be viewed as constructing the signal using translates of the Dirichlet kernel (2.16)—so the distinction between using the WKS (sinc) kernel and using the transform kernel becomes blurred. Incidentally the idea of using translates of some parent function to perform reconstruction is implicit in the radial basis approach (§2.4.3).

2.4.2 Using the transform kernel

Low-pass reconstruction

First let us find the reconstruction when there are N uniformly spaced samples x_n (intersample spacing $1/2W$) and the underlying signal is bandlimited to W. Take the interval $[-W, W]$ and identify its end-points. N equally-spaced frequencies f_m are marked off, starting at $f = 0$. The reconstruction is posed as a linear combination of the functions

$$B_m(t) = e^{2\pi i f_m t}.$$

As the time samples are uniformly spaced the weights can be obtained by the DFT (because the DFT is invertible) and $x^r(t)$ can be found explicitly as

$$x^r(t) = \sum_n x_n \frac{\sin a(t - t_n)}{N\binom{\sin}{\tan}\frac{1}{N}a(t - t_n)} \tag{2.16}$$

(with sin used when N is odd, tan when N is even). This may be thought of as the Lagrange interpolant for a function specified at N points evenly distributed round the unit circle. As $N \to \infty$ this tends towards the Shannon reconstruction. Note that the extension of this function, outside the observation interval, is periodic (period $N/2W$). A crucial point which will be used again is that the frequency spacing of the basis functions ($2W/N$) is the reciprocal of the observation time ($N \times 1/2W$).

Nonuniform sampling

Let us consider the same situation when the sampling is irregular. In that case we pose a model of the form

$$x(t_n) = \sum_{m=1}^{M} \lambda_m B_m(t_n) \qquad (1 \le n \le N), \qquad B_m(t) = e^{2\pi i f_m t} \tag{2.17}$$

In the previous section we argued that the spacing of the basis frequencies should be the reciprocal of the observation time T. This argument was based on regular sampling, but there are two reasons why it should be true in general. First, over an observation time T two signals $\exp(2\pi i f_1 t)$ and $\exp(2\pi i f_2 t)$ differ 'significantly' if and only if $|f_1 - f_2| > 1/T$ (in fact when $|f_1 - f_2|$ is a nonzero integer multiple of $1/T$ they are orthogonal). Secondly the reconstructed signal is composed of harmonics that are integer multiples of the frequency spacing δf, and will therefore be periodic with period $1/\delta f$; if $\delta f > 1/T$ then $1/\delta f < T$ and the reconstruction would have to repeat inside the observation window, whereas the underlying signal $x(t)$ could not be expected to do this.

Having established that $\delta f < 1/T$, the number of unknown parameters is $M = 2W/\delta f > 2WT$. For correct identification of these (by linear least-squares) we require $N > M$, and so $N/T > 2W$. As N/T is the average sampling rate, we have found that the average sampling rate must exceed the signal bandwidth[5]. This is in agreement with the point hinted at in §2.3(d).

Given that we have chosen the frequencies f_m to be equally spaced, we may write down the least-squares solution to the above problem:

$$\begin{aligned} \boldsymbol{\lambda} &= \mathbf{A}^{-1}\mathbf{b} \\ A_{rs} &= \sum_{n=1}^{N} e^{2\pi i(s-r)\delta f t_n} \\ b_r &= \sum_{n=1}^{N} x_n e^{-2\pi i r \delta f t_n} \end{aligned}$$

The matrix $\mathbf{A}$ is hermitian and Toeplitz and so the computational cost of inverting it is the square of the matrix dimension (by the Levinson recursion: see [128] for details of this) rather than the cube, which would be the cost if $\mathbf{A}$ were an arbitrary matrix. There are other methods based on the enlargement of a Toeplitz matrix to a circulant one (in which case inversion can be effected using the FFT) or conjugate-gradient methods. The idea can be extended to higher dimensions, whereupon the matrix assumes block-Toeplitz structure. A full discussion of these issues, and some examples of their application, is in the papers of the NUHAG group ([63, 64, 169, 207, 208] + references therein).

[5]Recall that we are defining the bandwidth B as the measure of the support of the Fourier transform.

Multiband signals

We can reconstruct a *multiband* signal in the same way, provided we know whereabouts in the frequency domain the bands lie. In each band the basis frequencies must be spaced by $\leq 1/T$, so there are $\geq B/(1/T)$ of them; for stable reconstruction, N must be greater than this, and so $N/T > B$ as before. Numerical simulation confirms that this train of thought is justified [141]. This is a famous result proved in a rather less rustic fashion by Landau [115, 116] and developed by Katsnelson [105]. Scoular & Fitzgerald [185] consider the reconstruction of real multiband signals with bands of equal bandwidth using periodic uniform sampling. Greitans [76] presents an example in which a real signal, with frequencies between 19Hz and 22Hz, is randomly sampled at an average sample rate $R = 17$Hz. One sees from this that $B = 6$Hz (the signal is real), which is substantially less than R: this explains why his methods worked well.

Periodic signals

Another situation in which equally-spaced frequency components arise is when the *underlying* signal is periodic (and of known period). In that case we can pose the same model as (2.17), except that now the basis frequencies are multiples of f_0 (the fundamental frequency). Again the problem reduces to a Toeplitz matrix inversion.

It is worth reflecting that there is an important information-theoretic difference between reconstructing a bandlimited signal using equispaced frequencies and reconstructing a periodic signal. In the first case the number of basis frequencies is $2WT \propto N$, and although the reconstruction is periodic the repetitions occur outside the observation window (because the period is T, the observation time). For signals of known period (the second case) the number of basis frequencies is $2W/f_0$ independently[6] of N, and the reconstruction repeats (several or many times) within the observation window. Defining the ratio

$$\lim_{T\to\infty} \frac{\text{Number of parameters } M(T)}{\text{Observation time } T}$$

as the information rate, we are led to the correct conclusion that the information rates are $2W$ for the first case and 0 for the second.

Sampling sets

The question that we are considering is this. Suppose that a signal x of bandlimit W is supplied. What are the admissible sampling sets, i.e. those sets of time instants that permit reconstruction? We have justified that an average sampling rate $> 2W$ is required, but have been a little vague on questions of sufficiency. It is quite a subtle question and there are several related questions of uniqueness and of stability. There are two ways in which one can examine it. First, we can consider the canonical

[6] Provided $2W/f_0 < N$ of course; otherwise the problem is underdetermined.

product with respect to the sample points,

$$S(t) = t^k \prod_n \left(1 - \frac{t}{t_n}\right)$$

(where k is the number of samples at the origin), and apply the method of contour integration to obtain a sampling expansion; we must then ask what conditions must be obeyed by the set $\{t_n\}$ for the Lagrange interpolant (2.13) to be correct.

The second approach is more subtle, and goes as follows. There is no difference between reconstructing the signal x and reconstructing its Fourier transform X. This can be uniquely determined only when $X(f)$ lies in the space generated by the functions $\exp(2\pi i f t_n)$. If we make the further assumption that $x \in L^2(\mathbb{R})$, so that $X(f) \in L^2(-W, W)$, then we are looking for a generating set (or Riesz basis, if we do not wish for overcompleteness) for $L^2(-W, W)$.

In this paragraph we shall for simplicity take $W = \frac{1}{2}$. It is clear that the set $\{t_n = n\}$ is admissible. It seems fair to assume that migration of the samples by a small amount should not affect the admissibility of $\{t_n\}$, and a famous result of Kadec [103] states that provided that each sample moves by less than some upper bound $< \frac{1}{4}$, admissibility is preserved. A generalisation by Avdonin [8] asserts that only the *average* displacement (in a well-defined sense) has to be less than $\frac{1}{4}$. This has been an active area of research for some time and some important results have been obtained by Beutler [15, 16]. He raises an interesting question about sampling in which only past samples are available, i.e. at $0, -1, -2, \ldots$. For such schemes an average sampling rate is not well-defined, and although this is a set of uniqueness (no other signal of bandlimit W agrees with $x(t)$ at those points) it is not a set of stable reconstruction (small changes in x produce large changes in the reconstructed signal). Recent work includes generalisations to $L^p(\mathbb{R})$ [86, 220], sampling at complex time instants [219], and multiband signals [105, 123, 124, 187]; see also [186]. Voss [220] provides a good overview.

2.4.3 More general reconstruction techniques

It is apparent from the above discussions that although reconstruction is possible using various techniques, those for general irregular sampling are not computationally straightforward and require large matrix operations.

This has led researchers to investigate approximate reconstruction of signals. In principle one can distinguish between two types of generalisation: (i) the underlying signal is still assumed to be bandlimited but the reconstruction does not have to be, and (ii) the underlying signal is assumed only to be 'slowly-varying'. Given that we wish a reconstruction algorithm to be stable to small unbandlimited perturbations, (ii) is more commonly supposed.

Marvasti ([143], §VII) gives a good overview of techniques. For example we have spline reconstruction ([3, 53] give introductory texts and Butzer [33, 37, 47] give examples of use in a sampling-theoretic context). Spline reconstruction is called a *local* technique, which means that at any given point only finitely many basis functions (or, equivalently, parameters) contribute to the reconstructed signal there. Simple Lagrange and sample-and-hold interpolators are also a popular choice.

Multidimensional signals (or functions, as they would normally be called) present a difficult problem: from the point of view of bandlimited functions one can use Fourier series, as discussed previously in this section and in [208]. A more flexible alternative, which has gained much popularity in recent years, is the use of radial basis functions (RBFs), to which Powell [163] provides an overview. The RBF approximant to a function from a metric space $(\mathcal{X}, d)$ to $\mathbf{R}$ is

$$t \mapsto \sum_{j=1}^{n} \lambda_j \psi(d(t, c_j)) \tag{2.18}$$

in which the c_j are known as 'centres', placed in $\mathcal{X}$, and ψ is some smooth nonlinear function such as a Gaussian ($r \mapsto e^{-\frac{1}{2}r^2}$), though $\ln(1+r^2)$, r^3, $\sqrt{1+r^2}$ and $r^2 \ln r$ may also be used. These have been extensively used in neural networks [200] and in chaotic time series prediction (§2.7). This is partly because one can place the centres in the domain of definition of the function. Recent work by the present author [137] suggests that RBFs outperform bandlimiting approximators when the function to be approximated is not smooth; in particular, he has shown that the delta-function approximant, rather than being a sinc function, is of the form $\gamma \sin(\pi t/\tau)/\sinh(\pi\gamma t/\tau)$, where τ is the intercentre spacing and γ relates to the 'radius' of the RBF. This of course decays very much more rapidly than the usual sinc function.

Now that the assumption of bandlimitedness has been relaxed, it is important to consider whether some minimum sampling rate is necessary for accurate reconstruction. It goes without saying that there is one, and these reconstruction techniques always work best at low frequencies and fail when the signal fluctuates significantly between the sample points. In rounding off this section, we point to the inevitable conclusion that, with the exception of the reconstruction of multiband signals with known bands, discussed in the previous section, no general technique is capable of accurate reconstruction when there are significant components above the Nyquist limit.

2.5 The spectrum

2.5.1 Introductory remarks

The power spectrum, or power spectral density (PSD), of a continuous-time signal is defined as the formal limit

$$P(f) = \lim_{T\to\infty} \frac{1}{2T} \left| \int_{-T}^{T} x(t) e^{-2\pi i f t} dt \right|^2$$

and it is used for assessing the spectral content of signals that have infinite energy (such as stationary stochastic processes, or signals generated by chaotic attactors). The power spectrum of a sampled signal, $P^s(f)$, is usually defined as the formal limit $N \to \infty$ of the periodogram function

$$P_N^s(f) = \frac{1}{N} \left| \sum_{n=1}^{N} x_n e^{-2\pi i f t_n} \right|^2 \text{ or } \frac{1}{2N} \left| \sum_{n=-N}^{N-1} x_n e^{-2\pi i f t_n} \right|^2 \tag{2.19}$$

though modifications to take irregular sampling into account can be made [129]. Given that they have been defined independently of each other, it is necessary to investigate how $P(f)$ and $P^s(f)$ are related. When the sampling is regular with rate R, $P^s(f)$ consists of equispaced translates of $P(f)$, spaced by R, in the same way that the Fourier transform behaves under sampling. In particular if $x(t)$ is bandlimited to W and $W < \frac{1}{2}R$ then the copies of $P(f)$ do not overlap and $P^s(f)$ determines $P(f)$ uniquely; but if $W > \frac{1}{2}R$ then this is no longer so.

When the sampling is irregular the position is more complicated but the conclusions are essentially the same, namely that $P^s(f)$ is the convolution of $P(f)$ with a function related to the sampling scheme [24]:

$$P^s = P * \mathrm{SC}. \tag{2.20}$$

By substituting a delta-function for $P(f)$ we can treat $\mathrm{SC}(f)$ as the spectrum of a sampled process taking the value 1 at each sampling instant, i.e. it is the observed spectrum of the constant process. Two interesting case are as follows.

With *additive-random* sampling, in which the intersample spacings $t_{n+1} - t_n$ are i.i.d. as $p(t)$, the function SC is obtained in terms of $\tilde{p}(f) = \int_0^\infty p(t)e^{-2\pi i f t}\,dt$ as

$$\mathrm{SC}(f) = \delta(\bar{t}f) + \mathrm{Cont}\,\mathrm{Re}\left(\frac{1+\tilde{p}(f)}{1-\tilde{p}(f)}\right) + \sum_{f_d \neq 0} h_d\,\delta(f - f_d). \tag{2.21}$$

(Here $\bar{t}$ is the mean intersample spacing. The operator 'Cont' signifies that the $\mathrm{Re}\,(\cdot)$ term is taken as continuous at the singular frequencies $f = f_d$ where $\tilde{p}(f) = 1$; in fact these must generate delta-function contributions, which have been explicitly brought out in the equation as the last term. Note that 0 is always a singular frequency, and it generates the first term. A full discussion is given in Appendix A.)

With *jittered sampling* the sampling instants vary from their mean sample positions by $t_n = (n + s_n)\bar{t}$, where s_n are i.i.d. as $q(t)$. Then SC is obtained in terms of $\tilde{q}(f) = \int_{-\infty}^{\infty} q(t)e^{-2\pi i f t}\,dt$ as

$$\mathrm{SC}(f) = 1 - |\tilde{q}(\bar{t}f)|^2 + |\tilde{q}(\bar{t}f)|^2 \sum_{n \neq 0} \delta(\bar{t}f - n). \tag{2.22}$$

It is interesting to compare these two sets of results. In general, SC for the additive-random scheme will contain a delta-function at the origin and a continuum (the second term in (2.21)) but no delta-function train. This is because $p(t)$ will only give rise to nonzero singular frequencies when it consists of delta-functions: however, such a possibility cannot be ignored, because that is exactly what happens for regular sampling[7]. On the other hand, SC for the jittered scheme will in general contain a delta-function train; the exceptions occur precisely when $\tilde{q}$ vanishes at all nonzero integers[8]. One might guess that, as the delta-function train in SC is responsible for aliasing, that additive-random sampling is better than jittered sampling at

[7] A union of periodic sampling schemes, with commensurate periods, is the only way in which this can occur. So regular sampling is a pathological case!

[8] For 100% uniform jitter, i.e. when $q(t) = 1$ for $|t| < \frac{1}{2}$ and 0 outside, this does occur, as $\tilde{q}$ is the sinc function.

suppressing aliases. This is true, and corroborated by some of the simulations in §4.6; see also [23].

An important point that is not emphasised in the literature is that the average of SC (over the real axis) is 1. This means that we cannot construct a sampling scheme in which only one copy of $P(f)$ comes through into $P^s(f)$; we must have unwanted copies as well and these two examples (of which the second is additive-random) give the extremes:

- Regular sampling: $\mathrm{SC}(f) = \sum_n \delta(\bar{t}f - n)$.
- Poisson sampling: $\tilde{p}(f) = (1 + 2\pi i f\bar{t})^{-1}$, and $\mathrm{SC}(f) = \delta(\bar{t}f) + 1$.

In the first case the unwanted copies give rise to aliasing; in the second they are smeared out to a uniform level. Let us give $P(f)$ a spike, or other pronounced feature, at frequency f_0. In the first case we cannot find f_0 uniquely from $P^s(f)$; in the second case we can. But in the second case the dynamic range is much reduced, and in a plot of $P^s(f)$ we would only be able to find the strongest features.

We have therefore identified three fundamental principles: irregular sampling suppresses aliases in a line spectrum; additive-random sampling does this better than jittered sampling; irregular sampling causes spectral smearing, reducing the dynamic range.

2.5.2 Practical difficulties

The periodogram function $P^s_N(f)$ suffers from three problems:

(a) Inconsistency. For stochastic signals, $P^s_N(f) \not\to P^s(f)$ as $N \to \infty$. This has been known for a long while (see e.g. [161]) and is a problem regardless of the sampling. A simple demonstration constructs FFTs of samples of white noise. As the sample length is increased the spectrum retains its jagged appearance rather than settling down to a smooth (constant) PSD estimate [128]. This can be mitigated by averaging periodograms of different length [148]; see also [26]§10.

(b) Leakage. This means simply that a spectral component at frequency f_0 appears at other frequencies too. Taking a finite amount of data corresponds to multiplication in the time domain by a gate-function, or convolving the Fourier transform with a sinc function, thereby giving each pronounced spectral feature a train of 'sidelobes'. By multiplying the data by a tapered 'window function' that decays to 0 at the edges, the sidelobes decay more rapidly (a corollary of the Riemann-Lebesgue lemma) but the resolution is reduced. If the data length is less than about 32 points, windowing gives hopeless results.

(c) Smearing. This is a consequence of irregular sampling and can be summarised quite neatly as: $P^s(f) \neq P(f)$. Even for a bandlimited signal, $P^s(f)$ and $P(f)$ do not agree on the Nyquist interval. This problem is often called leakage, as well as (b), but unlike the effects in (b) it cannot be cured by windowing. This is because a good estimate of $P^s(f)$, which is what the modifications in (a)

and (b) are trying to achieve, is not a good estimate of $P(f)$. The only way to get back to $P(f)$ is to deconvolve P^s with SC and although several methods have been proposed (see [25]) this is a very ill-conditioned problem [180] made worse by the fact that estimating $P^s(f)$ is difficult enough in the first place (problem (a)).

Scargle [180], who discusses these issues well (except (c), which he describes as leakage and attributes entirely to use of the 'raw' periodogram, rather than to the sampling), mentions in his introduction that despite all these problems the raw periodogram is still a good and quick method of detecting a sinusoid in noise, and that seems a sensible attitude to take. It is this that led him to discuss an alternative periodogram, which we mention next.

2.5.3 The Lomb spectrogram. Detection

An alternative spectral estimator, similar to the DFT, is the Lomb spectrogram [122]. This exploits the ability of the Fourier transform to detect a sinusoid in white additive noise. If we decompose $\{x_n\}$ thus,

$$x_n = \underbrace{A\cos(2\pi f t_n) + B\sin(2\pi f t_n)}_{s_n(f)} + u_n,$$

in which, for each f, the parameters A, B are adjusted so as to minimise $\sum_n |u_n|^2$, and define

$$P^s(f) \stackrel{\text{Lomb}}{=} \sum_n |s_n(f)|^2,$$

then we have a function whose peaks indicate the main frequencies. Indeed one can use this definition to give a statistical test for the presence of a sinusoid in white noise, because for each f, $P^s(f)$ is after appropriate normalisation exponentially distributed with unit mean [180]. Computational aspects are discussed in *Numerical Recipes* [165].

It is worth mentioning in general terms why the periodogram (or any of its relatives) is a useful device for detecting sinusoids. Suppose a complex-exponential, $Ae^{2\pi i f_0 t}$, is observed in white noise of variance σ^2. From a sample of size N form the spectrum; the height of the spike corresponding the sinusoid is NA^2 whereas the average level produced by the noise is at σ^2 and that produced by the spectral smearing of the sinusoid is somewhere between 0 and A^2 depending on the sampling scheme. The 'relative spike height' on a logarithmic scale is therefore

$$10\log_{10}\frac{NA^2}{\sigma^2(+A^2)}\ \text{dB}$$

so the acquisition of more data leads to a more pronounced spike. This effect is known as *coherent gain*; the periodogram owes its existence to it.

2.5.4 Alias-free sampling

The concept of alias-free sampling was introduced in a classic paper by Shapiro & Silverman [189]. In it they derived, for jittered and for additive-random sampling, results similar to those given in §2.5.1, working mainly with the sample autocorrelation,

$$r_m := \mathcal{E}_n x_n^* x_{n+m} \equiv \mathcal{E}^{t,x} x_n^* x_{n+m} = \int_0^\infty R(\tau) p_m(\tau) d\tau,$$

where p_m is the pdf of $(t_{n+m} - t_n)$. They also considered the difficult problem of determining $R(\cdot)$ from the r's. If no two distinct autocorrelation functions $R_{1,2}$ could generate the same set of r's the sampling scheme was said to be *alias-free*. The problem reduces to one of establishing whether the functionals $p_m(\tau)$ form a 'complete set', i.e. allow $R(\cdot)$ to be found. Remarkably one does not have to know the times at which the observations were made(!), as r_m depends only on the ordinates. A particular case is that of Poisson sampling, in which the $p_m(\tau)$ are (up to a few factors) the Laguerre functions and, coming as they do from a Sturm-Liouville problem, are complete[9]. So Poisson sampling is alias-free, and the above construction using Laguerre functions allows $R(\cdot)$ to be obtained. This property is exploited by Masry in the construction of a consistent spectral estimate for Poisson sampling [145, 146].

Shapiro & Silverman found that jittered sampling is not alias-free. They also found that additive-random sampling is alias-free if $\tilde{p}$, viewed as a function from $\mathbb{R}$ to $\mathbb{C}$, is injective ('one-to-one'), and is not alias-free if $\tilde{p}$, viewed as a function from the lower half-plane to $\mathbb{C}$, is not injective. As examples, the $\Gamma(\nu, \lambda)$ pdf is alias-free provided that $\lambda > 0$ and $\nu \leq 2$, whereas the rectangular pdf is never alias-free.

Since then their work has been reconsidered and extended by various authors. Strangely, nobody seems to have pointed out, in the additive-random case, that their equation ([189], p.236) for $P^s(f)$ with additive-random sampling is incorrect; one might also criticise that in the jittered case there is no explicit statement of SC(f). We have resolved these issues in Appendix A, and discussed the precise form of SC in the additive-random case. Beutler & Leneman [18, 19, 17] and Masry [144] have extended the definition of 'alias-free' so that one can define a sampling scheme as alias-free relative to a family of spectra. Masry [145, 147] points out that Shapiro & Silverman's definition of 'alias-free' is deficient in that it does not not necessarily imply that $P(f)$ can be consistently estimated from a finite set of samples; he strengthens the definition to suit. Bilinskis & Mikelsons [22] concentrate on signals with line spectra and argue that it does not matter too much whether a sampling scheme is (rigorously) alias-free, if it succeeds in suppressing aliases over a substantial spectral range. This is the basis of our discussion at the outset, and in Chapter 4 we shall feel free to use a variety of sampling schemes.

It cannot be emphasised too strongly that 'alias-free sampling', as defined here, has nothing to do with sampling rates and nothing to do with signal reconstruction; it is purely an exercise in identifying the spectrum. That is why no mention of Nyquist limits has been made; Poisson sampling, for example, is alias-free for all

[9] Specifically $\sum_n u_n(\tau) u_n(t) = \delta(t - \tau)$ where u_n are the normalised eigenfunctions.

(nonzero) sampling rates. In information-theoretic terms there is a great difference between knowing the spectrum and knowing the waveform: it is only for a very special class of signals—those consisting of sinusoids in observation noise—that the problems of spectral analysis and signal reconstruction become equivalent.

2.5.5 Final points

Of all the problems associated with spectral analysis of irregularly sampled data, the smearing problem is the most difficult one to solve and it has important consequences, not just for signal detection. The fact that a sinusoid no longer transforms to a simple spike means that there is no inverse Fourier transform available to us, and that we cannot go freely between time and frequency domains, as we are used to in classical digital signal processing (DSP). This has serious repercussions, one of which is that designing filters for irregularly sampled data has to be done in the time domain; as Marvasti points out, this requires a time-varying impulse response ([143], p.6), and Bilinskis states that filtering is one of the major unsolved problems in 'irregular DSP' ([23], §11).

2.6 Linear models

2.6.1 Introductory remarks

We have already seen that when the bandwidth of a signal is small compared with the sampling rate, the signal has, in effect, been oversampled and consequently is predictable, at least over a short range. This has been addressed in the sampling-theoretic literature in recent years [36, 39, 140]. The principal thrust of this research has been to construct one-sided sampling series, as opposed to the bi-infinite sum (2.4). Its main disadvantage is that the notion of bandlimitedness is not always very helpful, particularly in the case when the signal consists of several narrow-band components.

The statistical, or stochastic, approach is subtly different in that it associates the notion of linear prediction with that of modelling and *parametrisation*; the prediction coefficients, central to the description of a signal, give a tighter representation than a simple bandlimit. (If the fit is poor, the fitting procedure will say so.) The abundance of signals that are susceptible to parametric spectral analysis has led to a vast literature on linear prediction and spectral estimation, to which good introductions are [125, 107, 128]. Some specific examples are radar clutter (see e.g. [142]) and audio signals (see e.g. [74] and [77]§8).

The autoregressive model celebrates its seventieth anniversary this year but, judging by the continuing steady stream of papers, it is not showing its age. It was introduced by Yule [227] to find periodicities in the sunspot data, on the basis that a sinusoid $A\cos\Omega t + B\sin\Omega t$ is linearly predictable from two equispaced previous observations:

$$x_n = -a_1 x_{n-1} - x_{n-2}, \qquad a_1 = -2\cos\Omega\delta t$$

and that to account for prediction errors one needed an additive 'disturbance term'

on the RHS. After a further generalisation the autoregressive, or AR(p), model was born[10]:

$$x_n = -\sum_{j=1}^{p} a_j x_{n-j} + \varepsilon_n.$$

The term $\{\varepsilon_n\}$ is an uncorrelated Gaussian process with variance σ_ε^2.

Linear processes are usually analysed in the z-domain, with z defined by $z = e^{2\pi i f \delta t}$; this change of variable maps the real frequency axis to the unit circle in the z-plane. Let $A(z) := 1 + \sum_{j=1}^{p} a_j z^{-j}$. Then an AR process is generated by passing white noise into a system of transfer function $1/A(z)$. The zeros of A, (α_i) say, are called the *poles* of the model; for the process to be meaningful it must be stable, which means that the poles must be inside the unit circle. In the special case when all the poles tend toward the unit circle, and $\sigma_\varepsilon^2 \to 0$, the process becomes completely coherent, or harmonic, and consists of p tones at angular frequencies Ω_i given by $\alpha_i = \exp(i\Omega_i \delta t)$. In that case the prediction errors are zero; we call this the *coherent case.* The spectrum of an AR process is given by

$$P_{\mathrm{AR}}(z) = \frac{\sigma_\varepsilon^2}{A(z)A^*(1/z)} \overset{|z|=1}{=} \frac{\sigma_\varepsilon^2}{\left|1 + \sum_{j=1}^{p} a_j z^{-j}\right|^2}$$

which is continuous (on account of the fact that the model is assumed to hold for all time) and may have a large dynamic range irrespective of the size of the data set used to construct the AR coefficients. By contrast the periodogram (q.v.) is a polynomial in z, and only has well-defined features when the coherent integration time is large.

A further generalisation is the ARMA(p, q) model, in which a 'moving-average' term is introduced on the RHS:

$$x_n = -\sum_{j=1}^{p} a_j x_{n-j} + \sum_{j=0}^{q} b_j \varepsilon_{n-j}$$

This gives rise to a system function that contains zeros (as well as poles).

There are several excellent accounts of AR and ARMA modelling, for example [107, 125] and the texts [128] and [77].

2.6.2 Coefficient estimation and maximum-entropy

The coefficients can be estimated either from the sample autocorrelation function (Yule-Walker method) or by minimising the prediction error power defined as

$$E(\mathbf{x}, \mathbf{a}) = \frac{1}{2(N-p)} \sum_{n=p+1}^{N} \left\{ \left| x_n + \sum_{j=1}^{p} a_j x_{n-j} \right|^2 + \left| x_{n-p} + \sum_{j=1}^{p} a_j^* x_{n-p+j} \right|^2 \right\}$$

[10] Some authors correctly distinguish between the process, which is a sequence of random variables and written $X_n = \sum_{j=1}^{p} a_j X_{n-j} + \varepsilon_n$, and the observations, which are written in lower-case letters. However it is more usual, particularly in the engineering literature, to use the same symbol for both. This should not cause confusion.

with respect to the AR coefficients in an optimisation that can be either constrained (Burg) or unconstrained (Covariance). The constraint in the Burg method is the Levinson recursion, which relates the pth-order model $a_{1\ldots p}^{(p)}$ to the $(p-1)$th-order model $a_{1\ldots p-1}^{(p-1)}$ by a single coefficient ρ_p, the *reflection coefficient*:

$$a_j^{(p)} = a_j^{(p-1)} + \rho_p a_{p-j}^{(p-1)*} \qquad (1 \le j < p), \qquad a_p^{(p)} = \rho_p.$$

This is a byproduct of the Toeplitz matrix inversion in the Yule-Walker method; it guarantees that the model is stable whenever the reflection coefficients are all inside the unit circle[11]. The second $|\cdot|^2$ term is the backward prediction error energy, occurring because if $\{x_n\}$ obeys the AR model with coefficients a_j then its time-reverse obeys the model with coefficents a_j^*.

The above method is the maximum likelihood formulation (if one takes forward prediction errors only and assumes Gaussian driving noise), for one is in essence maximising the probability of observing the given dataset w.r.t. the AR parameters. In recent years the theory of *entropy* has led to new insights into the AR problem. The entropy of a process is a measure of its information content or 'disorder'. For a discretely sampled Gaussian process the entropy (more correctly the entropy per sample) is given in terms of the power spectrum as

$$\text{Entropy} = 2\delta t \int_{-1/2\delta t}^{1/2\delta t} \ln P(f)\, df$$

which (rather nicely[12]) is simply $\ln \sigma_\varepsilon^2$ for a stable AR process, independently of the AR parameters. The maximum-entropy approach to linear prediction seeks to find the PSD with maximum entropy (or the whitest spectrum) that agrees with the data as far as the first p autocorrelation lags. Solving by means of Lagrange multipliers, one finds that $P(f)$ must be the reciprocal of a polynomial in z. See [77]§7.2.4 and [107] for good discussions. The solution is then the same as that of the Yule-Walker method. In information-theoretic terms it corresponds to constructing the unknown autocorrelation moments in a way that 'makes fewest assumptions'.

An important issue that arises when fitting models is what order to use. It is not difficult to see that the quality of fit improves with the model order p, for the residual variance $\hat{\sigma}_\varepsilon^2[p]$ decreases. One therefore imposes a penalty on high-order models that is severe when the available data length is short. Two such methods are the Akaike Information Criterion and Minimum Description Length criteria :

$$\begin{aligned} \text{AIC}[p] &= N \ln \hat{\sigma}_\varepsilon^2[p] + 2p \\ \text{MDL}[p] &= N \ln \hat{\sigma}_\varepsilon^2[p] + (\ln N)p \end{aligned} \tag{2.23}$$

though there are others (see [128]). Unfortunately none of them works particularly well on short data sets [217], which is the field in which AR modelling has most to offer.

[11] Straightforward consequence of Rouché's theorem in complex analysis [205].

[12] Cepstrum lovers will recognise it as the zeroth cepstrum coefficient. The cepstrum of AR (and ARMA) processes is a particularly elegant function of the z-plane poles and zeros, carrying the implication that these estimators are fundamentally logarithmic.

2.6.3 Irregular sampling

As the AR model is a discrete one it is not surprising that the early work on irregular sampling concentrated on the missing data problem. There are three approaches, as follows. Let $\mathcal{K}$ be the set of indices n for which x_n is known.

- Method of Nuttall [155]. Here E is calculated only over the available prediction errors; the nth forward prediction error is available if and only if $n, \ldots, n-p \in \mathcal{K}$. The problem with this method is seen when the missing data points are scattered; in that case, the number of available prediction errors may be too small.

- Maximum-entropy method [153, 154] using the autocorrelation function (ACF). Here the autocorrelation lags $R_{xx}[k]$ are estimated from available pairs:

$$R_{xx}[k] = \frac{\sum' x_j^* x_{j+k}}{\sum' 1} \qquad \text{(Missing data problem)}$$

  in which the $'$ symbol means that the only values of j included are those for which both j and $j+k$ are in $\mathcal{K}$. The AR coefficients can then be found by maximum-entropy.

- Expectation-Maximisation (EM) algorithm [92, 191]. E is quadratic in $\mathbf{x}$ and in $\mathbf{a}$, so it can be minimised iteratively, w.r.t. the AR coefficients keeping the unknown data fixed, then w.r.t. the unknown data keeping the AR coefficients fixed.

In the special case when the ACF is known explicitly, there is the following option:

- Irregular ACF method [56]. Occasionally the ACF of the continuous-time process is given at irregular intervals, e.g. when the data come from an interferometer or autocorrelator. Then the maximum-entropy method is used to obtain the AR spectrum.

For more general sampling these methods cannot be used and the following have been suggested:

- Reconstruct, resample and use standard methods [149]. This is acceptable only when the signal is oversampled, as is often the case in geophysics.

- Continuous-time methods [101, 102]. Here the continuous-time all-pole model is posed:

$$\{\mathcal{L}y(t)\}dt = dB(t), \qquad \mathcal{L} \equiv \sum_{j=0}^{p} b_j \left(\frac{d}{dt}\right)^{p-j}.$$

  Defining $\underline{\mathrm{x}}(t) = [y(t) \ \cdots \ y^{(p-1)}(t)]^\top$ and $\underline{\mathrm{m}} = [1 \ 0 \cdots 0]$ we can write the above stochastic differential equation as

$$\frac{d}{dt}\underline{\mathrm{x}}(t) = \underline{\mathrm{B}} \cdot \underline{\mathrm{x}}(t) + \underline{\mathrm{e}}(t)$$

for an appropriate matrix $\underline{\mathrm{B}}$, and the observations as $y_n = \underline{\mathrm{m}} \cdot \underline{\mathrm{x}}(t_n)$. The Kalman filter (see e.g. [80]) is a predictor-corrector scheme that converges to an estimate of the state vector $\underline{\mathrm{x}}$ even though only one component is observed. Given a set of putative coefficents $\mathbf{b} = (b_j)$ one may thus derive a likelihood function $\mathrm{lik}(\mathbf{y}|\mathbf{b})$ and maximise it w.r.t. $\mathbf{b}$. In principle this method should produce very good ('optimal') results but in Jones' experience [101, 102] there were problems because the function $\mathbf{b} \mapsto \mathrm{lik}(\mathbf{y}|\mathbf{b})$ was 'ill-behaved', having a large number of local maxima. This is clearly at variance with the regular case, in which the prediction error function is quadratic in the unknown AR parameters.

- Generalised prediction error method [132, 134, 136, 139]. This is ours (see Chapter 4). Briefly the idea is to replace the prediction of x_n from p previous values with a generalised concept of prediction[13]:

$$\left|\sum_{j=0}^{p} r_j^n x_{n-j}\right|^2 \ll \left|\sum_{j=0}^{p} r_j^n\right|^2 \left|\sum_{j=0}^{p} x_{n-j}\right|^2$$

The LHS is the generalised prediction error energy. In the coherent case (q.v.) it can be arranged to be zero. It can also be made to coincide with the conventional definition when the sampling is regular. It was conceived for the missing data problem; subsequent work moved to the continuous-time AR model à la Jones, and then the two were combined. Although a nonlinear optimisation is required to find the model poles, the performance surface appears smooth and remarkably free from spurious minima.

2.6.4 Linear models and linear filtering

The AR and ARMA processes are closely linked to the theory of linear filtering. A finite impulse response (FIR) filter[14] with input $\{x\}$ and output $\{y\}$ is defined by

$$y_n = x_n + \sum_{j=1}^{p} a_j x_{n-j}$$

from which it is clear that passing an AR process with parameters (a_j) through an FIR filter with coefficients (a_j) gives white noise. The catch is that this filter is not the optimal filter for separating an AR process from an additive combination of it and another process. To see why, consider the case in which the AR process is a constant function ($p = 1$, $a_1 = -1$); the optimal filter for the removal of this is a notch filter at DC. But the FIR filter given by ($p = 1$, $a_1 = -1$) does not have a notch characteristic. To solve the problem properly one needs to write down the Wiener filter (which amounts to assuming Gaussian statistics and doing maximum-likelihood). Indeed, let $x(t)$ and $s(t)$ be uncorrelated signals, of known spectra $P_x(f)$

[13] By replacing $\ll$ with $\leq$ one would simply obtain the Cauchy-Schwarz inequality.

[14] See [150] for a good introduction to the practicalities of the subject.

and $P_s(f)$, and let their sum be given as $z(t)$. The Wiener filters for estimating x and s from z have, respectively, transfer functions

$$\frac{1}{1+P_s(f)/P_x(f)} \quad \text{and} \quad \frac{1}{1+P_x(f)/P_s(f)}$$

which are acausal and IIR in general. This is derived in many texts, e.g. [77, 104, 160]. What happens if $P_s(f)$ is unknown and $P_x(f)$ is known only up to a factor, though? This does not seem to have been discussed at all, and we show how to resolve the problem in Chapter 4.

When one tries to filter irregularly sampled data, the literature is even less helpful. Apart from the Savitzky-Golay smoothing filter (a simple device based on local polynomial fitting [179, 165]) and the 'top-hat' filter implicit in signal reconstruction (wherein frequencies under the Nyquist limit are passed and the others rejected), which are both specific types of filter, the only attempt to construct linear filters seems to have been made by Bilinskis and co-workers [23]. Their idea is to construct a 'notch-pass' filter by performing a discrete convolution with a sinusoid. (This is based on the elementary observation that in continuous time one would construct such a filter by convolving with a sinusoid.) There are several difficulties with this method: as an FIR filter, it requires a large number of taps to achieve a narrow passband; the transfer function is ill-defined on account of sampling irregularities; and one cannot construct stop-band filters. Chapter 4 will show how to deal with these problems and others, using the generalised prediction error approach.

2.7 Nonlinear models

There exists a large class of signals that are not well modelled using Fourier, or linear, methods, but are susceptible to nonlinear prediction. An artificial example is data from the Hénon map, $x_n = 1 - 1.4x_{n-1}^2 + 0.3x_{n-2}$, which is observed to be very broad-band and has a featureless power spectrum [184]. Indeed by examining the spectrum one would not be able to distinguish it from white noise, but of course the time series is (in principle, at least) perfectly predictable, unlike noise. For such signals, nonlinear prediction provides a much better method of attack than linear techniques: but how likely are such signals to occur? Suppose that the signal arose from observing a dynamical system

$$dy/dt = f(y) \qquad f : \mathcal{M} \to \mathcal{M}$$

at periodic intervals every τ (time units). Here $\mathcal{M}$ is a D-dimensional compact differential manifold. This is a structure that looks locally like $\mathbb{R}^D$, in the sense that a small open set in $\mathcal{M}$ is homeomorphic (i.e. in 1-1 correspondence via a continuous map) to an open set in $\mathbb{R}^D$. For example a torus is a compact differential manifold of dimension 2.

Suppose—and this gives the complications—that we do not observe $y(t) \in \mathcal{M}$ explicitly, but instead observe it through a smooth function $g : \mathcal{M} \to \mathbb{R}$. Also let $\phi : \mathcal{M} \to \mathcal{M}$ be the map that takes y to where the dynamics send y after time τ.

We may construct[15] the set of *delay-vectors*

$$\mathbb{E} = \left\{ [\, g\phi^{-1}y \quad g\phi^{-2}y \quad \cdots \quad g\phi^{-d}y \,]^{\mathsf{T}} : y \in \mathcal{M} \right\}$$

and give the mapping that performs this construction a label, Φ say. So

$$\Phi : y \mapsto [\, g\phi^{-1}y \quad g\phi^{-2}y \quad \cdots \quad g\phi^{-d}y \,]^{\mathsf{T}}.$$

We have a picture, in which the bottom branch has yet to be supplied:

$$\begin{array}{ccc} \mathcal{M} & \xrightarrow{\phi} & \mathcal{M} \\ \Phi\downarrow & & \Phi\downarrow \\ \mathbf{E} \subset \mathbb{R}^d & \xrightarrow{?} & \mathbf{E} \end{array} \tag{2.24}$$

We now employ the following theorem of differential topology [210, 90].

Theorem 1 *(Takens Embedding à la Huke) Let $\mathcal{M}$ be a compact manifold of dimension D. Let $\phi : \mathcal{M} \mapsto \mathcal{M}$ be a diffeomorphism satisfying the following constraints (which are generic):*

- *ϕ has a finite number of periodic points with period $\leq 2D$*
- *if y is a periodic point with period $k \leq 2m$ then the eigenvalues of ϕ^k at y are distinct.*

Then for all g in an open dense subset of $C^2(\mathcal{M}, \mathbb{R})$, the map $\Phi : \mathcal{M} \to \mathbb{R}^d$ given by

$$\Phi : y \mapsto [\, g\phi^{-1}y \quad g\phi^{-2}y \quad \cdots \quad g\phi^{-d}y \,]^{\mathsf{T}}$$

is an embedding if $d \geq 2D + 1$. □

A simple justification is that it is a generic property of D-dimensional manifolds that they do not intersect in $(2D + 1)$-dimensional space: in other words if they do intersect, an arbitrarily small perturbation in some direction will break the intersection. By saying that the constraints on ϕ are generic, we mean that if the constraints are not so for a particular choice of ϕ then they will be true for one arbitrarily close to it. We can think of a non-generic ϕ as being 'unlikely to occur'. As an example of the use of the terminology, a real $(n \times n)$ matrix is generically nonsingular, if we give the space of $(n \times n)$ matrices the Euclidean topology of $\mathbb{R}^{n^2}$; in fact the set of nonsingular $(n \times n)$ matrices is open dense in $\mathbb{R}^{n^2}$.

That Φ is an embedding means that it is smooth and invertible. As $\mathbb{E}$ is by construction the image of Φ, we have that Φ gives a smooth 1-1 correspondence (diffeomorphism) between $\mathcal{M}$ and $\mathbf{E}$. Define now the following map to complete the bottom branch of the diagram (2.24), marked by '?', and make the diagram commute:

$$K = \Phi\, \phi\, \Phi^{-1} : \mathbf{E} \to \mathbf{E}$$

[15] In composition of maps we shall not always put in the parentheses, so $g\phi^{-1}y$ is short for $g(\phi^{-1}(y))$.

which is a diffeomorphism because Φ and ϕ are. Obviously invertibility of Φ is very important: otherwise K is not well-defined. The map Φ would not be invertible if the embedding dimension d were too low.

Having established that a suitably well-behaved K exists, we can now apply it to time series prediction. Write

$$x_n = gy(t + n\tau), \qquad \underline{x}^n = [\, x_{n-1} \quad x_{n-2} \quad \cdots \quad x_{n-d} \,]^\top$$

so that $\{x_n\}$ form a time series and

$$K : \underline{x}^n \mapsto \underline{x}^{n+1}.$$

The 'top' component of K therefore takes $\underline{x}^n$ to x_n; write H for that function. Then

$$x_n = H(\underline{x}^n) \tag{2.25}$$

and the time series is predictable using the smooth function H.

From the dynamicist's point of view the importance of Takens' theorem is that it says that under generic conditions $\mathcal{M}$ and E are diffeomorphic, and time-discretised dynamics on $\mathcal{M}$ are manifested in E. From the practitioner's point of view the figure $d = 2D + 1$ is salient, as it is small enough to make practical the estimation of H from experimental observations. (Indeed d can often be made smaller: for the Lorenz and Rossler systems [212], in which $D = 3$, we find that $d = 4$ suffices, though the embedding theorem says that an embedding exists for $d \geq 7$.)

Consequently these methods have stimulated a vast range of applications including fluid dynamics, electronic engineering, biology, medicine and economics (for good overviews see e.g. [158, 106]). It has led both to a reexamination of old data sets and to the construction of new experiments, with the aim of detecting deterministic behaviour in time series previously thought to be random. The subject is often informally referred to as 'chaotic time series analysis' (for good overviews see e.g. [1, 199]).

The function H can be estimated from a time series by local or global functional approximation [60, 43, 61, 1]. A particularly attractive idea is the radial basis method, because H is not defined on $\mathbb{R}^d$ but rather on E which is a subset of it, and so one can choose centres in that region. In particular we may choose delay-vectors from the time series as centres, and the following approximation to H can be posed:

$$x_n \approx \hat{H}(\underline{x}^n) = \sum_{j=1}^{m} \lambda_j \psi(\|\underline{x}^n - \underline{x}^{c_j}\|).$$

The (c_j), which are just integers, index the centres.

In practice we must be able to choose the (c_j) and find the (λ_j). The latter problem is accomplished by linear least-squares, in which for a block of training data $(x_n)_1^N$ the following cost function is to be minimised:

$$\text{Error} = \sum_{n=d+1}^{N} \left(x_n - \sum_{j=1}^{m} \lambda_j \psi(\|\underline{x}^n - \underline{x}^{c_j}\|) \right)^2 = \|\mathbf{A}\boldsymbol{\lambda} - \mathbf{b}\|^2$$

where $\mathbf{A}$ is an $((N-d)\times m)$ matrix with entries $A_{ij}=\psi(\|\underline{x}^i-\underline{x}^{c_j}\|)$ and $\mathbf{b}$ is an $((N-d)\times 1)$ vector with entries $b_i=x_i$.

We wish to find the centres that have smallest effect on the error; the corresponding basis functions are then removed from the basis set (which corresponds to removing those columns from $\mathbf{A}$). This can be done independently of finding the 'best' weight-vector $\boldsymbol{\lambda}$, or selection and parameter estimation can be combined. The second approach is discussed in detail by Stark [200]. We shall discuss the ideas underpinning the first approach from the point of view of radial basis function fitting.

The 'independent' approach consists in choosing the centres first and then finding the (λ_j). Choosing a large number of centres at random from the training data is a good way of ensuring that the domain of the function to be fitted is well covered with centres. Many of the centres, however, will be too close to each other and this causes the matrix $\mathbf{A}$ to become ill-conditioned. A simple idea for eliminating this problem is to ensure that no two centres are less than some predetermined distance [195]. To evaluate 'closeness' in the context of two identical radially-symmetric[16] functions $f_{1,2}$ centred at $\underline{c}^{1,2}\in\mathbb{R}^d$, one may use as a yardstick the value $f_1(\underline{c}^2)$ $[=f_2(\underline{c}^1)]$. Any quantitative conclusion depends of course on the functions in question. From our experience it appears that for a Gaussian RBF $f:\underline{x}\mapsto e^{-\frac{1}{2}w\|\underline{x}-\underline{c}\|^2}$, for which $f(\underline{c})=1$, a centre should be rejected if the value of its associated basis function at any of the previously chosen centres exceeds 0.9. This figure of 0.9 is independent of w.

It is now a question of finding $\boldsymbol{\lambda}$. In principle one can solve the 'normal equations', $\mathbf{A}^\top\mathbf{A}\boldsymbol{\lambda}=\mathbf{A}^\top\mathbf{b}$, which give the solution to the minimisation of $\|\mathbf{A}\boldsymbol{\lambda}-\mathbf{b}\|^2$, by Cholesky decomposition of $\mathbf{A}^\top\mathbf{A}$. A preferable approach if the basis set is degenerate (or almost so) is singular-value decomposition of $\mathbf{A}$. This expresses $\mathbf{A}=\mathbf{QDP}^\top$ with $\mathbf{Q},\mathbf{P}$ orthogonal, and $\mathbf{D}$ diagonal with elements $\delta_1\geq\delta_2\geq\cdots\geq\delta_m\geq 0$ (m is the number of columns of $\mathbf{A}$). The normal equations can then be cast in the form $\mathbf{DP}^\top\boldsymbol{\lambda}=\mathbf{Q}^\top\mathbf{b}$. Now those elements of $\mathbf{P}^\top\boldsymbol{\lambda}$ corresponding to the smallest of the (δ_i) cannot be reliably found because errors in the 'data vector' $(\mathbf{Q}^\top\mathbf{b})$ are greatly amplified on division by the small δ_is. SVD sets these elements of $\mathbf{P}^\top\boldsymbol{\lambda}$ to zero. To determine which of the (δ_i) should be regarded as 'too small', a tolerance C is set by the user and the integer r (the *rank*) satisfying $\delta_r\geq C\delta_1\geq\delta_{r+1}$ is found. The $(r+1)$th, $(r+2)$th, ..., mth elements of $\mathbf{P}^\top\boldsymbol{\lambda}$ are then set to zero, while the others are obtained in the obvious way by dividing each element of $\mathbf{Q}^\top\mathbf{b}$ by its corresponding δ_i. Then the matrix $\mathbf{P}$ is applied to obtain $\boldsymbol{\lambda}$. As a rule C should reflect the accuracy of the data; if it is too small, the solution vector $\boldsymbol{\lambda}$ will be unstable to small perturbations in $\mathbf{b}$. We usually use a value in the range 10^{-4}–10^{-5} in these sorts of simulation.

Recursive schemes for updating the (λ_j), i.e. those suited to sequential data, are discussed in [199].

All this was on the subject of regular sampling. There is less information on how to construct nonlinear models when the sampling is irregular, apart from some interesting work by Sauer [175, 176, 177] on reconstruction from interspike intervals. In

[16] We mean that the functions are identical in all particulars except for where they are centred.

Chapter 5 we shall discuss methods of fitting nonlinear models to irregularly sampled data; this relies on a generalisation of Takens' theorem, due to Stark, Broomhead and coworkers [202, 198], that incorporates the intersample spacings.

2.8 Summary

We conclude by making some general observations on the general theory of signal sampling and techniques for analysing irregularly sampled data. The latter part of the discussion is slanted towards the development of the thesis.

- Signal reconstruction requires some sort of minimal (average) sampling rate; in the 'low-pass' case this is twice the highest frequency present and in the 'multiband' case it is the bandwidth (the measure of the support of $X(f)$). This is justifiable on information-theoretic grounds. Equivalently we can talk about a *Nyquist limit* (half the average sampling rate). If a signal is bandlimited to less than that frequency, it is *oversampled* and low-pass reconstruction is possible. If not, low-pass reconstruction will fail.

- Sometimes reconstruction is given by a sampling series. In that case further processing should be done in a way consistent with that sampling series rather than through an injudicious application of delta-functions. This raises some interesting questions about how to perform convolution, filtering, system identification, etc., from discrete samples.

- Signals with a few unknown narrow frequency components above the Nyquist limit present a difficult problem as, with the exception of multiband reconstruction techniques, all reconstruction relies on oversampling. And multiband reconstruction is not very helpful either, because that requires the spectrum support to be known. It is interesting because *the situation of unknown narrow frequency components is that to which regular sampling has most to offer as an anti-aliasing device*, and it is not well-served by existing techniques.

- Estimation of spectra is an important problem, made difficult because of the smearing associated with alias-free sampling schemes. However there is in principle no Nyquist rate associated with spectral estimation: an alias-free sampling scheme in principle permits correct identification of the underlying spectrum regardless of sampling rate. This distinguishes spectral estimation from reconstruction, as does the following observation. Testing the hypothesis that a signal has a pronounced frequency component, against the null hypothesis that it is white, can be done without reconstructing the signal; moreover, attempting reconstruction without knowledge of the spectrum is virtually guaranteed to give the wrong answer.

- The above discussion has identified a class of *multiband* signals, which consist of a few widely-spaced narrow frequency components at unknown locations. (Such signals abound in Doppler radars [14, 120].) The information content of such signals is low, on account of their low bandwidth, so by Nyquist-Landau

they can be sampled at a rather leisurely rate provided that the sampling is done irregularly. The problem then is to identify such signals and, since Fourier techniques are likely to be quite useful, circumvent the spectral smearing problem. Present algorithms for this require the components to be sinusoidal (zero bandwidth) [20, 69] so that they can be sequentially extracted by least-squares techniques as their frequencies are identified in the Fourier spectrum.

- It is clear that autoregressive modelling would be most helpful in this task, because the characteristic AR spectrum consists of a few narrow spikes. This should allow the construction of a wide class of bandpass and bandstop filters to allow extraction of narrow-band spectral components in the time domain. However at present these techniques are insufficiently well-developed.

- The concept of nonlinear prediction should be applicable to irregularly sampled signals, and would probably be the best method of analysing signals of nonlinear dynamical origin.

Chapter 3

Developments in sampling theory

This chapter is divided into three parts. In the first we explore the properties of the generalised translation operator for the K-transforms and use it to derive results on minimum-energy reconstruction; then linear filters and the discrete convolution product are discussed; then the results are extended to general Volterra operators and it is shown that a Volterra operator can be identified from input and output samples taken using the same sampling scheme (even though this superficially gives rise to aliasing ambiguities). In the second section we discuss the contour integration method and derive some bounds for the truncation error for Bessel sampling. The third section also deals with discrete sampling and some results are derived concerning the identifiability of signals from samples taken at integer points; the theorems proved are difficult using real or complex analysis but, remarkably, are quite straightforward if p-adic power series are used.

3.1 Developments on Kramer's sampling theorem

3.1.1 Minimum energy reconstruction formula

We consider K-transforms as discussed in §2.2 and the reconstruction of a function whose K-transform is finitely supported on the interval I. By Kramer's theorem such a signal could be represented as

$$x(t) = \sum_{n\in\mathbb{X}} x(\ell_n)\frac{\int_I K(u,t)K(u,\ell_n)^*\rho(u)\,du}{\int_I |K(u,\ell_n)|^2\rho(u)\,du} \equiv \sum_{n\in\mathbb{X}} \frac{x(\ell_n)}{\kappa[n]}\,\mathrm{ke}\,(t\ominus\ell_n)$$

which requires the samples to be taken at time instants $t = \ell_n$. Suppose however that the observation points t_n are arbitrary, so that samples x_n are given at time points t_n. A reconstruction is required. Noting that the function $\mathrm{ke}\,(t\ominus\tau)$ has K-transform $1_{u\in I}K(u,\tau)^*$ which is zero outside I, we expect to able to reconstruct x using a more general form,

$$x^r(t) = \sum_{n\in\mathbb{X}} \xi_n\,\mathrm{ke}\,(t\ominus\tau_n), \qquad x^r(t_n) = x_n$$

in which the τ_n can be chosen arbitrarily. The ξ_n are then no longer the samples, but depend on the t_n, τ_n and x_n through a matrix equation that is easily seen to be

$$\mathbf{K}\boldsymbol{\xi} = \mathbf{x}, \quad K_{mn} = \mathrm{ke}\,(t_m \ominus \tau_n).$$

We wish to choose a specific set of τ_n and, in common with the regular case, it can be shown that the solution giving rise to the reconstructed signal of minimum energy

$$\|x\|^2 = \int_I |X(u)|^2 \rho(u)\, du = \sum_{n \in \mathbf{X}} \frac{|x(\ell_n)|^2}{\kappa[n]}$$

is obtained by putting $\tau_n = t_n$; and in that case $\mathbf{K}$ will be hermitian.

The proof proceeds on the same lines as Chapter 2 §2.4. Consider two possible reconstructions, $x(t)$ and $x^\circ(t)$, defined by

$$\begin{aligned} x^\circ(t) &= \sum_n \xi_n^\circ \,\mathrm{ke}\,(t \ominus t_n) \\ x(t) &= \sum_n \xi_n \,\mathrm{ke}\,(t \ominus \tau_n) \end{aligned}$$

and both interpolating the observation points, so $x^\circ(t_m) = x(t_m) = x_m$. Then

$$\begin{aligned} \|x - x^\circ\|^2 - \|x\|^2 - \|x^\circ\|^2 &= -2\mathrm{Re} \sum_{r \in \mathbf{X}} \sum_{m,n} \frac{1}{\kappa[r]} \xi_m^{\circ *} \xi_n \,\mathrm{ke}\,(\ell_r \ominus t_m)^* \,\mathrm{ke}\,(\ell_r \ominus \tau_n) \\ &\overset{\heartsuit}{=} -2\mathrm{Re} \sum_{m,n} \xi_m^{\circ *} \xi_n \,\mathrm{ke}\,(t_m \ominus \tau_n) \\ &= -2\mathrm{Re} \sum_m \xi_m^{\circ *} x_m \\ &= -2\mathrm{Re} \sum_{m,n} \xi_m^{\circ *} \xi_n^\circ \,\mathrm{ke}\,(t_m \ominus t_n) \\ &\overset{\heartsuit}{=} -2\mathrm{Re} \sum_{r \in \mathbf{X}} \sum_{m,n} \frac{1}{\kappa[r]} \xi_m^{\circ *} \xi_n^\circ \,\mathrm{ke}\,(\ell_r \ominus t_m)^* \,\mathrm{ke}\,(\ell_r \ominus t_n) \\ &= -2\|x^\circ\|^2. \end{aligned}$$

Therefore

$$0 \le \|x - x^\circ\|^2 = \|x\|^2 - \|x^\circ\|^2$$

and x° is the minimum-norm reconstruction. We have only to justify the step labelled ($\heartsuit$) above, namely the identity

$$\sum_{r \in \mathbf{X}} \frac{1}{\kappa[r]} \,\mathrm{ke}\,(\ell_r \ominus t_1)^* \,\mathrm{ke}\,(\ell_r \ominus t_2) = \mathrm{ke}\,(t_1 \ominus t_2)$$

which is a discretised and generalised version of the 'Hardy integral' discussed in §2.4. This result is a trivial consequence of Kramer's sampling theorem. Indeed, consider a function $y(t) = \mathrm{ke}\,(t \ominus t_2)$, and expand it from samples $\{t = \ell_r : r \in \mathbf{X}\}$:

$$y(t) = \mathrm{ke}\,(t \ominus t_2) = \sum_{r \in \mathbf{X}} \frac{1}{\kappa[r]} \,\mathrm{ke}\,(\ell_r \ominus t_2) \,\mathrm{ke}\,(t \ominus \ell_r)$$

Now put $t = t_1$, and note that $\mathrm{ke}(t_1 \ominus \ell_r) \equiv \mathrm{ke}(\ell_r \ominus t_1)^*$.

This is a very general method of proof not unknown in mathematical physics (for example it underpins the proof of the Minimum Dissipation Theorem in fluid dynamics, which states that of all velocity fields with the same boundary conditions the one with the smallest dissipation is the Stokes flow). More generally, these ideas stem from the following algebraic observation: if V is a vector space and U is a subspace of it, and we are to find the smallest element in a given coset $v + U$ of V/U, then that element is $x^\circ \in v + U$ satisfying $x \in v + U \Rightarrow \langle x, x^\circ \rangle = \langle x^\circ, x^\circ \rangle$. When the result is applied, U is the space of functions that vanish 'on the boundary' (in our case, this means signals that vanish at the sample points). Performing the minimisation using Lagrange multipliers is less elegant.

3.1.2 Linear filtering

We consider again a transform-limited signal $x(t)$, with transform $X(u)$ that is assumed to vanish for $u \notin I$. In (2.2) we stated that a linear filter can be regarded as a multiplication in the transform domain:

$$y(t) = (x \circledast f)(t) = \int_I X(u)F(u)K(u,t)\rho(u)\,du. \tag{3.1}$$

Because $X(u)$ vanishes outside I, we may as well assume that $F(u)$ does too (this is rather an important point). Then we may expand X and F as the DKTs (2.8) of their corresponding time samples:

$$y(t) = \sum_{r,s} \frac{x[r]f[s]}{\kappa[r]\kappa[s]} \int_I K(u,\ell_r)^* K(u,\ell_s)^* K(u,t)\rho(u)\,du.$$

Upon sampling we obtain the nth sample $y^s[n] = y(\ell_n)$ as

$$\begin{aligned} y^s[n] &= \sum_{r,s} Q_{nrs}x[r]f[s] \\ Q_{nrs} &:= \frac{1}{\kappa[r]\kappa[s]} \int_I K(u,\ell_r)^* K(u,\ell_s)^* K(u,\ell_n)\rho(u)du. \end{aligned} \tag{3.2}$$

If in the discrete time domain we implement the filter by

$$y^d[n] = \sum_{r,s} Q_{nrs}x[r]f[s] \tag{3.3}$$

then we have a commuting diagram

$$\begin{array}{ccc} x(t) & \xrightarrow{\mathcal{V}^c} & y(t) \\ \updownarrow & & \updownarrow \\ x[n] & \xrightarrow{\mathcal{V}^d} & y[n] \end{array}$$

in which $\mathcal{V}^c$ is (3.1) and $\mathcal{V}^d$ is the discrete convolution (3.3).

We may note that in the Fourier case Q_{nrs} is 1 if $n = r + s$ and 0 if not, so (3.3) reduces to a normal convolution of the samples of x and f.

3.1.3 Convolution products. Example

The formula (3.3) might well be referred to as the discretisation of the convolution product, but that name has already been given to the following result (see e.g. [98]):

$$(f \circledast g)(t) = \sum_n \frac{f(\ell_n) g(t \ominus \ell_n)}{\kappa_n} \tag{3.4}$$

for transform-limited f and g. The proof is straightforward, for by definition the LHS is

$$\int_I F(u) G(u) K(u,t) \rho(u)\, du$$

which on expanding $F(u)$ using the DKT evaluates to

$$\int_I G(u) K(u,t) \rho(u)\, du \sum_n \frac{K(u,\ell_n)^* f(\ell_n)}{\kappa_n};$$

now the integral is performed and the result drops out. Alternatively we can use the time-discretised Plancherel identity (2.9).

However we must remember that $g(t \ominus \ell_n)$ is not a sample of g, i.e. it is not the same as $g(t - \ell_n)$. Equation (3.4) does not, therefore, give a direct expression for the continuous-time convolution in terms of the samples. It is therefore only a partial discretisation; such nomenclature, if adopted, would be quite good because the DKT was applied only to $F(u)$ in the above derivation, whereas to obtain our full discretisation (3.3) we applied the DKT to both $F(u)$ and $X(u)$ in (3.1).

As an example let us consider some (very nearly) bandlimited functions that have closed-form J_0-Hankel transforms, namely the following class:

$$\alpha^{-1}\lambda \left(\frac{\alpha}{\sqrt{\alpha^2+t^2}}\right)^{\lambda+1} P_\lambda\left(\frac{\alpha}{\sqrt{\alpha^2+t^2}}\right) \leftrightarrow \frac{1}{\Gamma(\lambda)} \alpha^\lambda u^{\lambda-1} e^{-\alpha u} \tag{3.5}$$

where P_λ is the Legendre function, $\operatorname{Re}\lambda > -1$ and $\operatorname{Re}\alpha > 0$. (See [75], §6.621, and footnote[1]). Suppose an input signal $x(t)$ and a filter function $f(t)$ have parameters $\lambda_{x,f}$ etc.; then it is clear from the form of the Hankel transform that their convolution $y(t)$ will have parameters $\lambda_y = \lambda_x + \lambda_f - 1$, $\alpha_y = \alpha_x + \alpha_f$. For a specific test the following have been chosen:

Signal	λ	α
x	20	50
f	30	50
y	49	100

[1] Proof of (3.5). Write down the J_0-Hankel transform of the RHS. Replace the transform kernel $J_0(ut)$ with its integral representation, $\int_0^{2\pi} \exp(iut\cos\phi) d\phi/2\pi$. Do the u-integration. Disregarding a few numerical factors one is left with $\int_0^{2\pi} (\cos\theta \pm i \sin\theta \cos\phi)^{-\lambda-1} d\phi/2\pi = P_\lambda(\cos\theta)$, the integral representation of the Legendre function, with $\cos\theta = \alpha/\sqrt{\alpha^2+t^2}$. See [75] §8.411/§8.711.

The bandlimit has been selected as $b = 1$ (see §2.2.1), i.e. $X(u)$ and $F(u)$ are assumed to vanish for $u > 1$. The left-hand diagrams below show the time signals and the right-hand diagrams their transforms.

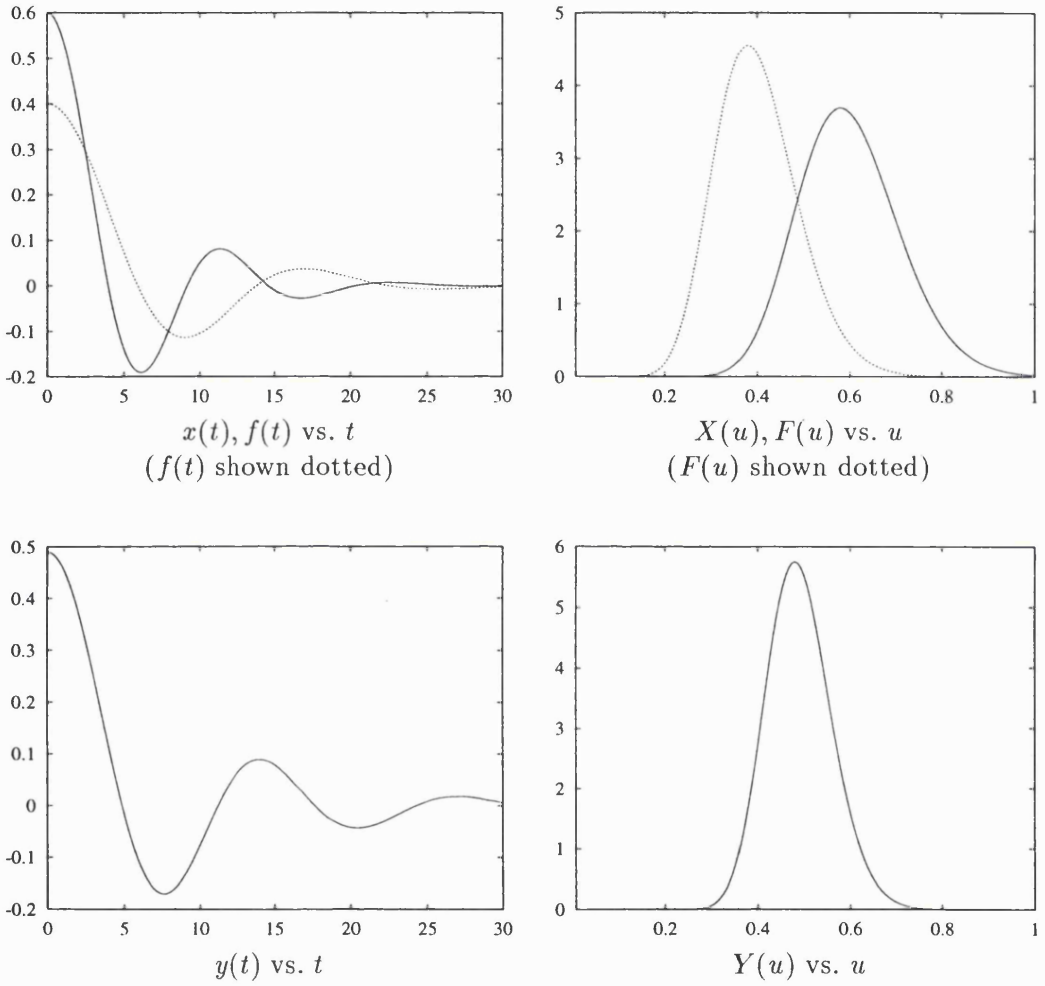


Figure 3.1. Test functions and their J_0-Hankel transforms.

The functions $x(t)$ and $f(t)$ were sampled at the zeros of $J_0(t)$, i.e. at $t = j_{0,n}$, for $1 \leq n \leq 10$. Next the discrete convolution (3.3) was employed to find the samples $y^d[n]$; of course, the sum had to be truncated ($1 \leq r, s \leq 10$) as there were only ten samples. Finally, the 'true' continuous-time waveform $y(t)$ was sampled and its samples, $y^s[n]$, compared with the calculated samples $y^d[n]$. The results are not quite the same, because two approximations were made: (i) assuming x and f to be transform limited, when they are not, and (ii) truncating the convolution sum. One can see from the plots of the signals and their transforms that the truncation error is small and that the aliasing error is minuscule. The next figure shows the extent to which y^s and y^d differ.

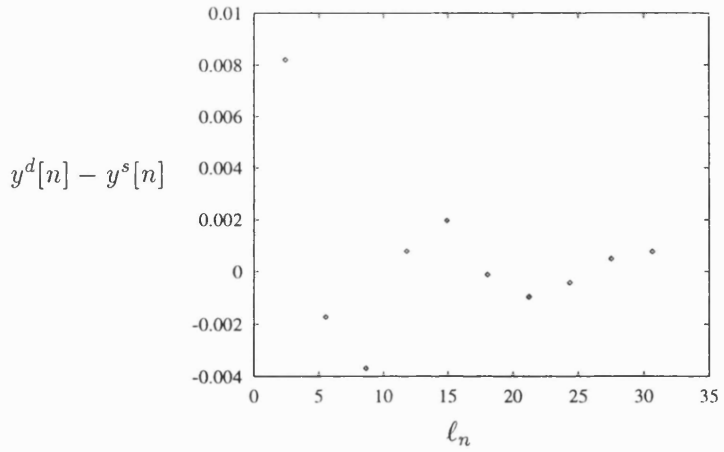


Figure 3.2. Error in truncated convolution product.

The error is consistently $\approx 2\%$ of $y(t)$, which seems to be a reasonable result in view of the short data record.

3.1.4 Nonlinear operators; the Volterra operator

We now consider a more general form of filtering, again for bandlimited inputs. The filter will now be nonlinear, though, and the diagram that we wish to establish lacks symmetry because $y(t)$ is no longer transform-limited:

$$\begin{array}{ccc} x(t) & \xrightarrow{\mathcal{V}^c} & y(t) \\ \updownarrow & & \downarrow \\ x[n] & \xrightarrow{\mathcal{V}^d} & y[n] \end{array}$$

The filter that we shall consider is, in the first instance, a quadratic Volterra operator. (The higher-order case follows directly.) This is an operator capable of introducing nonlinearity and memory, and in continuous time it is given by

$$y(t) = \iint h(\tau_1, \tau_2) x(t - \tau_1) x(t - \tau_2)\, d\tau_1\, d\tau_2$$

in the classical (Fourier) case. By writing h, x in terms of their Fourier transforms, we have

$$y(t) = \iint H(u_1, u_2) X(u_1) X(u_2) e^{2\pi i u_1 t} e^{2\pi i u_2 t} du_1\, du_2$$

General discussions on identification and correction of nonlinear distortion can be found in a series of publications by Tsimbinos & Lever [214, 215, 216]; the derivation we present here is a faster, more transparent, and more general version of Frank's proof [70]. Frank carries out his working in the frequency domain, which means that to see what happens when the output is sampled (the right-hand branch of the diagram) he needs to invoke the Poisson summation formula, complicating the proof.

Let us now develop this second expression in the same way as before, generalising first to the K-transform:

$$y(t) = \int_I \int_I H(u_1, u_2) X(u_1) X(u_2) K(u_1, t) K(u_2, t) \rho(u_1) \rho(u_2)\, du_1\, du_2$$

and then using the DKT to go back to the time domain:

$$\begin{aligned} y(t) = & \sum_{r_1, r_2, s_1, s_2} h[s_1, s_2] x[r_1] x[r_2] \times \\ & \int_I \int_I K(u_1, \ell_{s_1})^* K(u_2, \ell_{s_2})^* K(u_1, \ell_{r_1})^* K(u_2, \ell_{r_2})^* K(u_1, t) K(u_2, t) \\ & \frac{\rho(u_1)\rho(u_2)\, du_1\, du_2}{\kappa[s_1]\kappa[s_2]\kappa[r_1]\kappa[r_2]}. \end{aligned}$$

Notice as before that because $X(u)$ vanishes outside I, we may as well assume $H(\cdot, \cdot)$ to vanish outisde $I \times I$. That is why we have been able to use the DKT to expand it in terms of a two-dimensional array of time samples. Sampling each side and recalling the definition of the tensor Q :

$$y^s[n] = \sum_{r_1, r_2, s_1, s_2} Q_{n r_1 s_1} Q_{n r_2 s_2} h[s_1, s_2] x[r_1] x[r_2]$$

thereby coinciding with the 'obvious' discrete definition.

We can now write down the results for the higher-order Volterra operators. The general Volterra operator is given in continuous time by

$$y(t) = \sum_{m=0}^{\infty} \int_{I^m} H(\mathbf{u}) X(u_1) \ldots X(u_m) \rho(\mathbf{u})\, du_1 \ldots du_m$$

(where $\rho(\mathbf{u}) := \rho(u_1) \cdots \rho(u_m)$) and in discrete time by

$$y[n] = \sum_{m=0}^{\infty} \sum_{\mathbf{r}, \mathbf{s} \in \mathbb{Z}^m} Q_{n r_1 s_1} \cdots Q_{n r_m s_m} h^m[\mathbf{s}] x[r_1] \cdots x[r_m].$$

In each case m is the order of the nonlinearity and h^m its corresponding Volterra kernel.

3.1.5 Discussion

We have shown that a continuous-time Volterra operator induces a discrete-time operator in a natural way provided that:

- the input is transform-limited;
- the input may be reconstructed from samples via the symmetric K-transform reconstruction formula;
- the output and input are sampled using the same sampling scheme.

For the regular case, this just means that input and output are sampled simultaneously and at a rate exceeding the input Nyquist rate. One might expect that to identify a continuous-time operator would require the input to be sampled at its Nyquist rate and the output at *its* Nyquist rate (which might not even be finite). But that is not the case—a point that has been appreciated for some time, but not fully understood until recently (hence Tsimbinos' extensive work [214, 215, 216] and Frank's paper [70]).

What does this tell us about signal reconstruction? It is apparent that if a signal $y(t)$ has been obtained from a known bandlimited signal $x(t)$ by an unknown Volterra-type nonlinear distortion then the continuous-time waveform y can be reconstructed from samples taken at the Nyquist rate of x, which will generally be smaller than that of y. The reconstruction would be effected as follows. From the discrete samples (x^d, y^d) we identify the discrete Volterra operator $\mathcal{V}^d$, and having found its kernel we apply the integral transform to obtain $\mathcal{V}^c$. Then $\mathcal{V}^c$ is applied to $x(t)$ and the result is $y(t)$. We conclude that reconstruction of a signal not bandlimited to W has taken place from samples taken every $1/2W$. This is true, but extra information has been used, and that extra information is contained in x. Not every unbandlimited signal arises as a distortion of a bandlimited one. This is common sense, and information-theoretically it is clear that a distortion cannot increase the information rate even if it does increase the bandwidth, so the fact that reconstruction of y is possible at the input Nyquist rate is not surprising after all.

A related issue which should not be confused with the theory presented here is the *correction* (not identification) of nonlinear distortion. Consider the following problem. A signal $x(t)$ is bandlimited to W. It is then passed through an invertible memoryless nonlinearity, and then through a bandlimiter (bandlimiting to W). If the nonlinearity is known but the input is not, does the bandlimited output determine the input signal? The answer is 'yes' [114, 117, 173, 174] and the proof is based on the Contraction Mapping Theorem. The extension of this result to nonlinearities with memory is more delicate, because a Volterra operator does not necessarily have an inverse (for example it could simply be a linear filter). We suspect however that the case of the instantaneous nonlinearity should generalise to the symmetric K-transforms.

3.2 Truncation error bounds for Bessel sampling

The Shannon, Kramer-Weiss and other sampling expansions involve infinitely many terms. In practice one can only take finitely many, thereby giving rise to a truncation error; it is useful to have a convenient upper bound for this. The work in this section could be viewed as giving a truncation error bound for the Bessel sampling expansion *as obtained by the Kramer-Weiss sampling theorem*, but we shall not be using that theorem; we only use contour integration (§2.3). An elegant upper bound for the contour integration kernel allows a neat truncation error bound to be derived.

3.2.1 Derivations

As in §2.3 the objective is to obtain a sampling expansion by applying the residue theorem to the integral

$$I_{\square} = \frac{1}{2\pi \mathrm{i}} \oint \frac{x(\tau)}{(\tau - t)} \frac{S(t)}{S(\tau)} d\tau \,, \tag{3.6}$$

taken over the box contour (Figure 2.1) used in §2.3, and a truncation error bound by estimating the contributions from the left- and right-hand sides of the contour. This requires a lower bound for $|S(\tau)|$ on each of those sides. The assumptions used here are that:

- The signal $x(t)$ is real on $\mathbf{R}$ and bounded by M.
- The samples are taken at the zeros of the function $S(t) = J_0(bt)$.
- For $\tau \in \mathbf{C}$ $x(\tau)$ is of exponential-type $rb < b$.

Then as discussed in §2.3 we have $|x(\tau)| \leq M \cosh rbv$ ($v = \operatorname{Im} \tau$).

We can use the residue theorem to write down the following sampling expansion,

$$I_{\square} = \epsilon_T(t) = x(t) - \left(\sum_{n=-N}^{-1} + \sum_{n=1}^{N} \right) \frac{x(j_{0,n}/b)}{J_0'(j_{0,n})} \frac{J_0(bt)}{bt - j_{0,n}} \,, \tag{3.7}$$

which is simply the Lagrange interpolant. If we are to make any progress at all towards a truncation error bound, we must find lower bounds (or workable approximations) for $S(\tau) = J_0(b\tau)$ on each section of the contour. To estimate the contribution from the upper and lower sections we can use an asymptotic lower bound for J_0, which is that for any $\delta > 0$

$$|J_0(u \pm \mathrm{i}V)| > e^{(1-\delta)V}, \quad V \to \infty$$

Choosing $\delta = \frac{1}{2}(1 - r)$ we have

$$|x(\tau)/J_0(b\tau)| < Me^{rbV} \Big/ e^{(1-\delta)bV} = Me^{-\delta bV}$$

which ensures that these contributions vanish as $V \to \infty$. Next we consider the sides. We now use the asymptotic approximation to the Bessel function, which in fact gives a very good approximation to $|J_0(z)|$ on the line[2] $\mathrm{Re}\, z = j'_{0,n}$:

$$|J_0(z)| \approx \sqrt{\frac{2}{\pi|z|}} \cosh(\mathrm{Im}\, z) \qquad \mathrm{Re}\, z = j'_{0,n} \quad (n \neq 0) \tag{3.8}$$

The validity of such an approximation is confirmed by the following experiment, in which the following two functions,

$$M_1(z) = |J_0(z)|, \qquad M_2(z) = \sqrt{\frac{2}{\pi|z|}} \cosh(\mathrm{Im}\, z)\,,$$

were computed along the contour given by $\mathrm{Re}\, z = j'_{0,n}$. The percentage approximation error for M_1 by M_2,

$$\frac{M_2 - M_1}{M_1} \times 100\%,$$

was plotted. The results are shown here for $n = 1, 2, 3$.

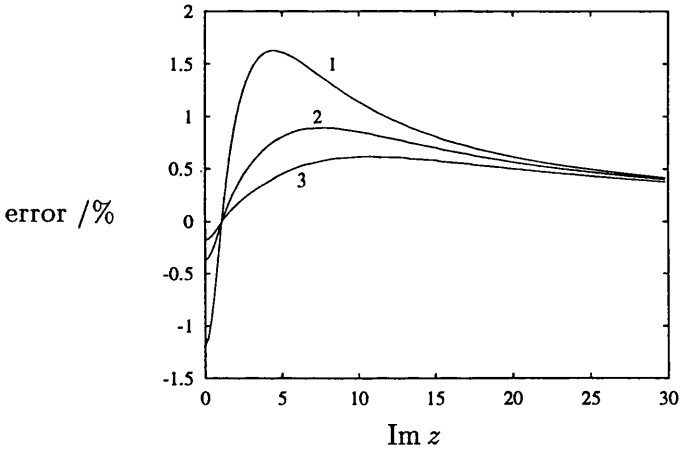


Figure 3.3. Error analysis for (3.8).

Having established that (3.8) is in fact an excellent approximation (improving as $n \to \infty$), we bound the side contributions by

$$\frac{M|J_0(bt)|}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\frac{|b\tau|^{1/2}}{|\tau - t|} \right) \frac{\cosh rbv}{\cosh bv} dv\,, \qquad \tau = \pm j'_{0,N}/b + \mathrm{i}v$$

and it is not difficult to show that, provided t is real, the parenthesised term is maximised on the real axis, taking the value $bj'^{1/2}_{0,N}/|j'_{0,N} - bt|$. Replacing it with this constant value and performing the integration gives the upper bound

$$|\epsilon_T(t)| \leq \frac{M\sqrt{\pi}|J_0(bt)|}{2\sqrt{2}\cos(r\pi/2)} \left\{ \frac{j'^{1/2}_{0,N}}{|bt + j'_{0,N}|} + \frac{j'^{1/2}_{0,N}}{|bt - j'_{0,N}|} \right\}. \tag{3.9}$$

[2]By convention $j_{0,n}$ $(n \neq 0)$ is the nth zero of J_0, and $j'_{0,n}$ is the nth zero of J'_0.

If different numbers of samples are taken on each side of $t = 0$, the variable N in the braced expression needs to be altered; the left- and right-hand terms come, respectively, from the truncation in the left- and right-hand summations in (3.7).

Remember that we have only assumed that $x(t)$ is an EFET bounded on the positive real axis. This has two advantages: first, it leads to an estimate for $|x(\tau)|$ that does not require knowledge of the Hankel transform of x, and secondly, $x(t)$ does not have to be even. Having said that, we can consider two special cases, namely when $x(t)$ is even or odd, and the latter allows us to improve our error estimate.

Even $x(t)$

We combine the integrals along the left- and right-hand sides of the contour to obtain

$$\epsilon_T(t) = \frac{J_0(bt)}{2\pi} \int_{-\infty}^{\infty} \frac{2\tau}{(\tau^2 - t^2)} \frac{x(\tau)}{J_0(b\tau)}\, dv, \qquad \tau = j'_{0,N}/b + \mathrm{i}v$$

and then estimate its magnitude:

$$|\epsilon_T(t)| \leq \frac{M|J_0(bt)|}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\frac{b^{1/2}|\tau|^{3/2}}{|\tau^2 - t^2|} \right) \frac{\cosh rbv}{\cosh bv}\, dv, \qquad \tau = j'_{0,N}/b + \mathrm{i}v$$

Again the $(\cdot)$ term is maximal when $v = 0$, so we simplify the integral as before and arrive at the following expansion and bound (which is essentially (3.9)):

$$\left| x(t) - \sum_{n=1}^{N} \frac{x(j_{0,n}/b)}{J'_0(j_{0,n})} \frac{2j_{0,n}J_0(bt)}{(b^2t^2 - j_{0,n}^2)} \right| \leq \frac{M\sqrt{\pi} j_{0,N}'^{3/2} |J_0(bt)|}{\sqrt{2}\cos(r\pi/2)|j_{0,N}'^2 - b^2t^2|} \tag{3.10}$$

Odd $x(t)$

Combining the integrals along the left- and right-hand sides of the contour as before, we obtain

$$\epsilon_T(t) = \frac{J_0(bt)}{2\pi} \int_{-\infty}^{\infty} \frac{2t}{(\tau^2 - t^2)} \frac{f(\tau)}{J_0(b\tau)}\, dv, \qquad \tau = j'_{0,N}/b + \mathrm{i}v$$

and then estimate its magnitude:

$$|\epsilon_T(t)| \leq \frac{M|tJ_0(bt)|}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\frac{|b\tau|^{1/2}}{|\tau^2 - t^2|} \right) \frac{\cosh rbv}{\cosh bv}\, dv, \qquad \tau = j'_{0,N}/b + \mathrm{i}v$$

Again the $(\cdot)$ term is maximal when $v = 0$, so we simplify the integral as before and arrive at the following expansion and bound:

$$\left| x(t) - \sum_{n=1}^{N} \frac{x(j_{0,n}/b)}{J'_0(j_{0,n})} \frac{2btJ_0(bt)}{(b^2t^2 - j_{0,n}^2)} \right| \leq \frac{M\sqrt{\pi} j_{0,N}'^{1/2} |btJ_0(bt)|}{\sqrt{2}\cos(r\pi/2)|j_{0,N}'^2 - b^2t^2|} \tag{3.11}$$

The error bound is 0 at the origin (despite the lack of a sample there), which is what it ought to be if x is known to be odd.

3.2.2 Simulations

The simplest examples of a function $x(t)$ satisfying the requirements enunciated at the beginning of this section are the sine and cosine. Then

$$\cos \omega t = \sum_{n=1}^{\infty} \frac{\cos(\omega j_{0,n}/b)}{J_0'(j_{0,n})} \frac{2 j_{0,n} J_0(bt)}{b^2 t^2 - j_{0,n}^2}$$

$$\sin \omega t = \sum_{n=1}^{\infty} \frac{\sin(\omega j_{0,n}/b)}{J_0'(j_{0,n})} \frac{2bt J_0(bt)}{b^2 t^2 - j_{0,n}^2}$$

with truncation errors given by (3.10) and (3.11) with $M = 1$. Of course, these results are valid only for $r = \omega/b < 1$.

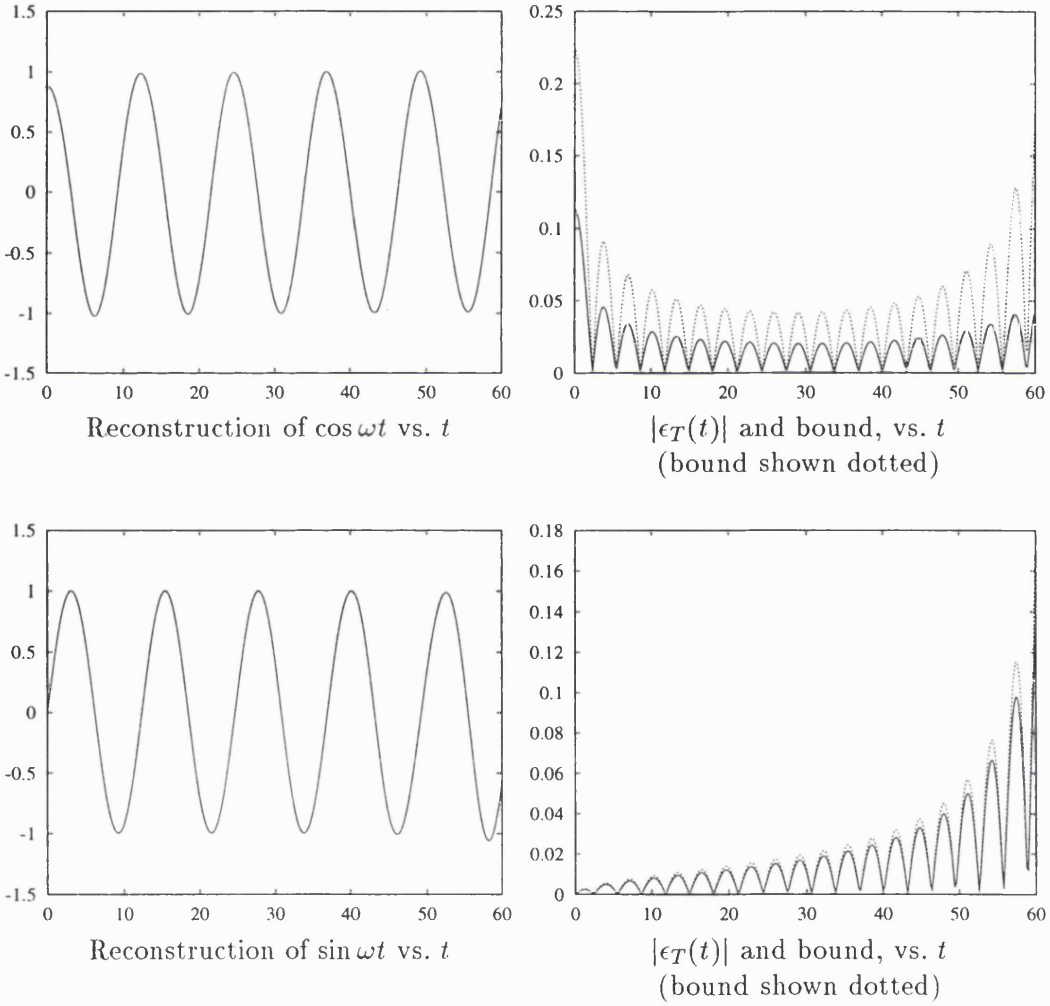


Figure 3.4. Truncation errors in the reconstruction $\cos \omega t$ and $\sin \omega t$ from samples taken at the zeros of $J_0(\cdot)$. Here $\omega = 0.51$.

3.2.3 Discussion

It is apparent that the truncation error bounds work well and are fairly easy to calculate. (In particular the truncation error tends to 0 as $N \to \infty$, a vital property of any sampling series.) The key step is the tight lower bound for the Bessel function, which is valid only along the lines $\operatorname{Re} z = j'_{0,n}$; by producing a bound of the form 'envelope $\times$ cosh(Im z)' we can follow the calculations through just as in the regular case, and produce similar results. Notice that the form of the lower bound is rather similar to the *upper* bound derived by Duffin & Schaeffer [57]. (For the Bessel function this would be $|J_0(z)| \leq \cosh(\operatorname{Im} z)$.) This is only to be expected. We only need a lower bound of $|J_0(z)|$ in regions of the complex plane where $|J_0(z)|$, viewed as a function of Re z, is locally maximal (this allows us to get a tight bound on the truncation error); and a tight lower bound for local maxima of a function will necessarily coincide with a tight upper bound for that function. So, rather paradoxically, in searching for lower bounds for use in this problem it is sensible to consider the form of the upper bound. In fact the general form of the lower bound is observed for other types of sampling, e.g. paired sampling [138] in which $S(t) = \cos at - \cos\beta$. Similar results apply to the sampling expansions of higher-order Bessel functions, in which $S(t) = (bt)^{-\nu} J_\nu(bt)$.

The fact that $x : \mathbb{C} \to \mathbb{C}$ needs only to be a bounded EFET is useful because it is more flexible than having to deal with the Hankel transform. For example the sine and cosine, which fulfil our requirements, do not have measurable J_0-Hankel transforms; they therefore do not meet the conditions of Jerri [100, 98][3] or Rawn [170] both of which require a bound to be placed on the Hankel transform of x. A useful spin-off of the results given here is that a self-truncating sampling series can be derived without any fuss. Going through the motions, if x has exponential-type $rb < b$ and $q = 1 - r$ then

$$x(t)\left[\operatorname{sinc}\frac{qb}{m\pi}(t'-t)\right]^m$$

is bounded on $\mathbf{R}$ and of exponential-type b. So it obeys our requirements and can be expanded as a sampling series in which the sampling instants are the zeros of $J_0(bt)$. Putting $t' = t$ gives

$$x(t) = \sum_{n\neq 0} x\left(\frac{j_{0,n}}{b}\right)\left[\operatorname{sinc}\frac{q}{m\pi}(bt - j_{0,n})\right]^m \frac{J_0(bt)}{J'_0(j_{0,n})(bt - j_{0,n})}, \tag{3.12}$$

which enjoys more rapid convergence than the series (3.7).

[3] Jerri's result has been published but not with proof:

$$|\epsilon_T(t)| \leq 2K\sqrt{(2/\pi)\ln(1-2r)^{-1}}\,|J_0(bt)/J_0(j'_{0,N})|\left(|j'_{0,N}+t|^{-1} + |j'_{0,N} - bt|^{-1}\right)$$

where

$$K = \int_0^r w^2 X(w)^2 dw.$$

This is a bizarre formula. For a start there seems no reason why K should be finite, and the mysterious $\ln(\cdot)$ term explodes when $r = \frac{1}{2}$, without good reason.

3.3 Ultrametric sampling theory

3.3.1 Basic ideas and a sketch theorem

This section discusses discrete sampling theory in a new way and gives, amongst other things, a rather nice result on the identifiability of sinusoids from samples taken at integer points. To start with let us consider in general terms the current relationship between discrete sampling and continuous sampling theorems.

The WKS and Kramer-Weiss sampling theorems make the obvious connection between discrete and continuous signals, using an assumption of transform limitedness. In §3.1 these have been used to derive consistent methods for processing discrete data. A continuous-time model may be used to describe *properties* of the samples. For example the autoregressive model is essentially discrete and its classical form says nothing about the evolution of the signal in between the samples; however as discussed in §2.6 and Chapter 4 it can be viewed as a discretely-sampled continuous-time stochastic process ('diffusion'). One may also consider the spectrum, and derive the spectrum of samples from the spectrum of an underlying stationary process. As a different type of example, work on nonlinear methods has shown that continuous-time dynamics, when observed periodically with an appropriate ('generic') observation function, induce discrete-time dynamics in the time series of observations (§2.7). In all that has been done there is a link between the samples and a supposed underlying waveform that may be real-valued, be complex-valued, or evolve on a manifold.

The work in this section breaks that link. Consider the following question:

> A discrete signal x^d from some family $\mathcal{C}$ is given at points $n \in \mathsf{L} \subset \mathbb{Z}$. Do the known samples $(x_n)_{n\in\mathsf{L}}$ cause x^d to be uniquely determined at each *integer* point?

If they do, we shall say that L determines x (over $\mathcal{C}$). For example if $\mathcal{C}$ is the set of rational functions regular at all integer points then L determines x over $\mathcal{C}$ if and only if L has infinitely many elements. If $\mathcal{C}$ is the set of functions bandlimited to W then L determines x over $\mathcal{C}$ if and only if the density of points in L exceeds $2W$ (see §2.4.2). Now for the crunch: what if $\mathcal{C}$ is the set of functions generated by finite numbers of harmonics? What can we say about them? In general terms if $\mathcal{C}$ is closed under addition and subtraction, as it has been in these three examples, then the question boils down to distinguishing between the zero function (which is in $\mathcal{C}$) from any other function in $\mathcal{C}$ on the basis of the samples on L.

Sketch Theorem 1 *Let $x(t)$ be a sum of harmonics[4] sampled for $t \in \mathsf{L} \subset \mathbb{Z}$. If, and only if, L has no periodic gaps, then the samples on $\mathbb{Z}$ are uniquely determined by those on L.*

We are *not* saying that the (real) continuous-time waveform is uniquely determined by the samples on L, for that would contradict some basis notions of aliasing. What we are saying is that the remaining samples (those in $\mathbb{Z}$ but not in L) are determined by the known samples (those on L).

[4] $x : t \mapsto \sum_{i=1}^{q} A_i \exp(\mathrm{i}\omega_i t)$, for constants $A_i \in \mathbb{C}$, $\omega_i \in \mathbb{R}$.

We can suggest an application of these results in array processing. The direction of arrival of a 'target' generating a sinusoidal signal is determined by performing signal processing analogous to the identification of the frequency of a tone from time-domain samples: the temporal frequency corresponds to the spatial angle. What we are saying is that not all the elements of an infinite equispaced array are necessary to perform this task, and moreover that from that array one may delete arbitrarily many elements provided that the thinned array does not contain periodic gaps. The general idea is well-known in the array processing community (see e.g. [87]) but an explicit identifiability theorem such as the Sketch Theorem would appear to be a novel contribution to the subject. The rest of this chapter is devoted to the exposition of the relevant mathematics.

So how are we to go about proving the Sketch Theorem? At first sight one would say that complex variable theory should do the trick. However the Sketch Theorem is more closely connected with the subject of exponential Diophantine equations [190]. Diophantine equations are the subject of questions such as 'Find all integer or rational solutions to [equation]', and they usually require number-theoretic techniques. The Diophantine theory of elliptic curves [45] (an elliptic curve is one of the form '$y^2 =$ cubic in x') is one such example, and much progress has been made on it using local fields $\mathbb{Q}_p$. These fields are somewhat analogous to $\mathbb{R}$ in that they are obtained from $\mathbb{Q}$ by completing (= filling in the gaps to make topologically complete) with respect to a *valuation* ('measure of size'); indeed $\mathbb{R}$ is sometimes written $\mathbb{Q}_\infty$. The valuation that is used in constructing the $\mathbb{Q}_p$ is the p-adic valuation which is defined number-theoretically. The p-adic valuation has some properties that are not paralleled in $\mathbb{R}$, and so Analysis in $\mathbb{Q}_p$, usually called *Ultrametric Analysis*, is rather different from Real Analysis. The reader is advised to consult Appendix B.

So, rather than considering the existence of an underlying signal defined on $\mathbb{R}$ or on $\mathbb{C}$, we are going to consider one defined[5] on $\mathbb{Q}_p$, and why not? After all one can do Analysis on $\mathbb{Q}_p$, and $\mathbb{Q}_p$ contains the integers $\mathbb{Z}$. The following diagram summarises the position, with the top line representing the conventional view in signal processing, and the right-hand side representing ours:

$$\begin{array}{ccc} (x^c : \mathbb{R} \to \mathbb{C}) & \underset{\text{reconstruct}}{\overset{\text{sample}}{\rightleftarrows}} & (x^d : \mathbb{Z} \to \mathbb{C}) \\ & & \uparrow ? \\ & & (x^p : \mathbb{Q}_p \to \mathbb{Q}_p) \end{array} \tag{3.13}$$

This is not the first time that p-adic methods have been used in signal processing, as number-theoretic and p-adic transforms are being used to design convolution and filtering operations (see e.g. [108]); but apart from a well-known paper on the generalisation of the sampling theorem to abstract groups [111], the paths of algebra and sampling theory cross only rarely.

[5] In fact we only need to define it on $\mathbb{Z}_p$, the ring of valuation-integers of $\mathbb{Q}_p$, i.e. the elements $y \in \mathbb{Q}_p$ satisfying $|y|_p \leq 1$.

3.3.2 Ultrametric interpolation

If (3.13) is to be of any use, we must be able to make some headway with the following problem:

Given a sequence $x = (x_n)_{n\in\mathbb{Z}}$, find a function g convergent on $\mathbb{Z}_p$ such that whenever ν is a (rational) integer n, $g(\nu) = x_n$.

We have defined $\mathcal{W}_p$ to be the space of such sequences (in Appendix B have referred to them as functions $\mathbf{Z} \to \mathbb{Q}_p$; clearly this is equivalent). We shall exclude $p = 2$ from now on. Then from Appendix B Lemma 10 we have that

$$\alpha \in \mathrm{Pos}(\mathbb{Q}_p) \Rightarrow (n \mapsto \alpha^n) \in \mathcal{W}_p$$

and also that for $\alpha \in \mathbb{Q}_p$,

$$|\alpha|_p = 1,\ (p-1) \mid m \Rightarrow \alpha^m \in \mathrm{Pos}(\mathbb{Q}_p).$$

So for $\alpha \in \mathbb{Q}_p$,

$$|\alpha|_p = 1,\ (p-1) \mid m \Rightarrow (n \mapsto \alpha^{mn}) \in \mathcal{W}_p.$$

So although the sequence (α^n) is not generally in $\mathcal{W}_p$, its $p-1$ subsequences, obtained by starting at r (say) and going along in steps of $p-1$, are in $\mathcal{W}_p$. To take this into account we enlarge $\mathcal{W}_p$, as follows.

Definition 1 *Let $x : \mathbf{Z} \to \mathbb{Q}_p$ and define the function $x^{[r]}$ by*

$$x^{[r]} : n \mapsto x_{r+n(p-1)}.$$

If, for each r between 1 and $p-1$, we have $x^{[r]} \in \mathcal{W}_p$, then we say that $x \in \mathcal{W}_p^+$. More compactly

$$x \in \mathcal{W}_p^+ \iff \left(1 \le r \le p-1 \Rightarrow (n \mapsto x_{r+n(p-1)}) \in \mathcal{W}_p\right).$$

Proposition 1 *If $p \neq 2$, $\alpha \in \mathbb{Q}_p$, $|\alpha|_p = 1$, then $(n \mapsto \alpha^n) \in \mathcal{W}_p^+$.*

For $x^{[r]} : n \mapsto \alpha^r.\alpha^{n(p-1)}$; clearly $(n \mapsto \alpha^{n(p-1)}) \in \mathcal{W}_p$, so the result follows. □

Theorem 1 *If $p \neq 2$, $\alpha_{ij} \in \mathbb{Q}_p$, $|\alpha_{ij}|_p = 1$, then*

$$\left(n \mapsto \sum_i q_i(n)\alpha_{i1}^n\alpha_{i2}^{n^2}\cdots\alpha_{ie(i)}^{n^{e(i)}}\right) \in \mathcal{W}_p^+$$

in which the sum is finite and the q_i are polynomials defined over $\mathbb{Q}_p$.

Proof. As $\mathcal{W}_p^+$ is a ring we have only to show that $(n \mapsto \alpha^{n^e})$ is in $\mathcal{W}_p^+$, where e is a (rational) integer, and α obeys the hypotheses of the theorem. To do this, let $x : n \mapsto \alpha^{n^e}$. Then

$$x^{[r]} : n \mapsto \alpha^{[r+(p-1)n]^e}.$$

Now

$$\alpha^{[r+(p-1)n]^e} = \alpha^{r^e}\alpha^{mn}$$

with m a multiple of $p-1$: so $x^{[r]} \in \mathcal{W}_p$ as required. □

3.3.3 Sampling lattices and functions in $\mathcal{W}_p^+$

Definition 2 *Λ_1 is the set[6] of subsets $\mathsf{L} \subset \mathbb{Z}$ that have the following property: for every integer r and every nonzero integer m there exists an element of L equivalent to $r \pmod m$.*

Essentially Λ_1 is the set of subsets of $\mathbb{Z}$ that do not have periodic gaps.

Proposition 2 *Λ_1 contains sampling sets of arbitarily low density.*

To see this, choose an irrational real γ and a subset J of $[0,1)$ of Lebesgue measure $|J| > 0$. Define for real t the number $[t]_1$ as the element in $[0,1)$ that differs from t by an integer. Define $\mathsf{L} \subset \mathbb{Z}$ by

$$n \in \mathsf{L} \Longleftrightarrow [n\gamma]_1 \in J.$$

By the density (Kronecker's) theorem for irrationals (see [79], Thm 439), $\mathsf{L} \in \Lambda_1$. And L has density $|J|$. □

Proposition 3 *Let $\mathsf{L} \in \Lambda_1$. For each r and each nonzero m there are in fact infinitely many elements $l \in \mathsf{L}$ satisfying $l \equiv r \pmod m$.*

Suppose that only s elements of L are equivalent to $r \pmod m$. By hypothesis there are elements of L equivalent $\pmod{(s+1)m}$ to $r, r+m, \ldots, r+sm$ respectively. These are distinct, there are $s+1$ of them, and they are each equivalent to $r \pmod m$, a contradiction. □

Theorem 2 *If $x \in \mathcal{W}_p^+$ for some[7] p, and if $\mathsf{L} \in \Lambda_1$, then L determines x.*

Proof. Suppose that there are two functions in $\mathcal{W}_p^+$ that agree on L. Let x be their difference; then x is in $\mathcal{W}_p^+$ and is zero on L. For each r between 1 and $p-1$, $x^{[r]} : n \mapsto x_{r+n(p-1)}$ vanishes for infinitely many $n \in \mathsf{L}$ (by Proposition 3) and so by Strassmann's theorem (Appendix B, Theorem 11) it is identically zero (viewed as a function on $\mathbb{Z}_p$). Therefore x vanishes at each integer, and that is all we need to show. □

Theorem 3 *Suppose that $x : \mathbb{Z} \to \mathbb{C}$ is of the form*

$$x : t \mapsto \sum_i q_i(t) e^{p_i(t)},$$

in which the sum is finite and p_i, q_i are polynomials. Any $\mathsf{L} \in \Lambda_1$ determines x.

[6] In previous versions of this work the condition on L was written $\varprojlim \bigcup_{l \in \mathsf{L}} l + m\mathbb{Z} = \mathbb{Z}$. This is equivalent.

[7] One does not have to know p; it suffices to know that some p will do the job.

Proof. Let the coefficients of the polynomials be p_{ij}, q_{ij}. Write

$$K = \mathbb{Q}(\{q_{ij}, e^{p_{ij}}, \text{ each } i,j\}).$$

Then K is finitely generated and by the Embedding theorem (Appendix B, Theorem 7) can be embedded in $\mathbb{Q}_p$ for infinitely many choices of p. We therefore choose $p \neq 2$, and define

$$\begin{aligned} y &: \mathbb{Z} \to \mathbb{Q}_p \\ y &: n \mapsto \sum_i q_i(n) e^{p_i(n)}. \end{aligned}$$

By Theorem 1 $y \in \mathcal{W}_p^+$. Now apply Theorem 2. □

Theorem 4 *(Converse to Theorem 3.) If* $\mathbf{L} \notin \Lambda_1$ *then there are two signals of the form*

$$x : t \mapsto \sum_i q_i(t) e^{p_i(t)}, \qquad \left.\begin{matrix} q_i = 1 \\ p_i = \textit{linear} \end{matrix}\right\} \forall i$$

that agree on $\mathbf{L}$ *but not on* $\mathbf{Z}$.

Proof. By hypothesis there exist m, r such that no element of $\mathbf{L}$ is equivalent to $r \pmod m$. Let

$$x : n \mapsto \frac{1}{m} \sum_{k=1}^{m} e^{2\pi i k(n-r)/m}.$$

Then

$$x(n) = \begin{cases} 0, & m \not| (n-r) \\ 1, & m \mid (n-r) \end{cases}.$$

So $x(\mathbf{L}) = \{0\}$, which means that x and the zero function agree on $\mathbf{L}$ but not on $\mathbf{Z}$, and that completes the proof. □

3.3.4 Discussion

Note that Theorems 3 and 4 together give quite a strong result. $\mathbf{L} \in \Lambda_1$ is, by the simple considerations of Theorem 4, obviously the minimum possible condition for determination of exponential signals, and it admits sampling sets of arbitrarily low density. The fact that it is a condition sufficient to determine not only these but also a much larger class $\mathcal{C}$, including exponentials, polynomials and exponentials of polynomials, is remarkable. Of interest is that $\mathcal{C}$ is a differential ring[8].

[8] Structure closed under addition, subtraction, multiplication and differentiation.

Chapter 4

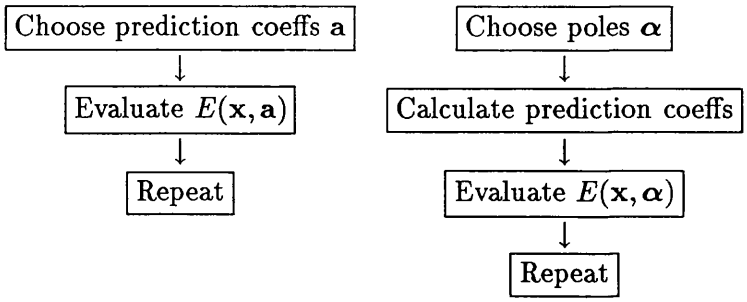
Linear models

In this chapter we shall discuss time-domain techniques for autoregressive (AR) modelling and AR-based signal separation using a generalised prediction error technique in which the prediction coefficients depend on the observation intervals and on the underlying poles. Spectral estimation is effected by minimising a certain error energy function with respect to the poles; this coincides with the Covariance method in the regular case. The resulting spectral estimate is quite subtle and has the following significance. Suppose that the observed data come from a continuous-time harmonic or autoregressive process $\mathcal{A}$. Then the discrete-time AR model that our method produces is the same as that obtained by fitting a conventional AR model to a regularly-sampled data record from $\mathcal{A}$. Such an approach requires a normalising frequency, or virtual sampling rate, which we call $1/\tau$. The construction of the error energy function arises from the interplay between the discrete and continuous-time models.

In separation schemes we show that the time-domain removal of features with known spectra is able to allow identification of smaller features made invisible by the spectral 'smearing effect' associated with irregular sampling.

4.1 Generalised prediction errors

In this first section we shall introduce the concept of generalised prediction for linear models. The objective is to define time-varying prediction coefficients, dependent on the sampling instants and on the underlying model poles, that generalise the regular case in a natural way. The estimation problem can be summarised as follows. In the regular case the poles do not enter the calculations explicitly, for we can just optimise with respect to the AR coefficients as the left-hand diagram on the next page suggests: the error function E is calculated and the prediction coefficients chosen to minimise it. In fact the optimisation is linear, so there is a closed-form solution and there is no need to go through this iterative process (though adaptive methods do). In the irregular case the poles are fixed, but not the coefficients, and this explains the extra step; further, the optimisation is nonlinear. Note that the optimisation with respect to the *poles* is nonlinear whether or not the sampling is regular.



Classical and generalised AR fitting.

4.1.1 Introductory remarks

Given that the prediction coefficients are time-varying, we establish how they are obtained from the poles. We see immediately that in the case of regular sampling the AR coefficients a_j are determined from the poles α_i by the matrix equation

$$\begin{bmatrix} \alpha_1^p & \alpha_1^{p-1} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_p^p & \alpha_p^{p-1} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \mathbf{0}$$

and the 'normalisation condition': $a_0 = 1$. The forward prediction errors are

$$\mathfrak{f}_n(\mathbf{y};\boldsymbol{\alpha}) = \sum_{j=0}^{p} a_j y_{n-j}. \tag{4.1}$$

At this point it is convenient to make some notational definitions. For a set of p complex numbers $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^p$ we define $A(z) = \prod_{i=1}^p 1 - \alpha_i/z$ and

$$\begin{aligned} N(\boldsymbol{\alpha}) &:= \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} A(z)A^*(1/z)\frac{dz}{z} \\ \frac{1}{C(\boldsymbol{\alpha})} &:= \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} \frac{1}{A(z)A^*(1/z)}\frac{dz}{z} \end{aligned}$$

System-theoretically these are, respectively, the power gains observed by putting white noise through the FIR filter with zeros α_i and the IIR filter with poles α_i. In terms of the coefficients a_j and the reflection coefficients ρ_i :

$$N(\boldsymbol{\alpha}) = 1 + \sum_{j=1}^{p} |a_j|^2 \quad \text{and} \quad C(\boldsymbol{\alpha}) = \prod_{i=1}^{p} 1 - |\rho_i|^2.$$

which give closed-form expressions for these quantities (expand $\prod_i(1 - \alpha_i/z)$ to get the coefficients a_j; this gives $N(\boldsymbol{\alpha})$, and the downward Levinson recursion (see §2.6.2) gives $C(\boldsymbol{\alpha})$).

4.1.2 Derivation for the coherent case

When the data consist of p pure tones at angular frequencies ω_i, the z-plane model poles are at $\alpha_i = e^{i\omega_i \delta t}$ and the prediction errors are all zero (independently of the amplitudes and phases of the tones). For irregular sampling let us define $\alpha_i = e^{i\omega_i \tau}$ in which τ is reasonably arbitrary; then to make the generalised forward prediction errors,

$$\mathfrak{f}_n(\mathbf{y};\boldsymbol{\alpha}) = \sum_{j=0}^{p} r_j^n y_{n-j}, \tag{4.2}$$

zero independently of the amplitudes and phases of the tones, we require

$$\begin{bmatrix} \alpha_1^{(t_n - t_{n-p})/\tau} & \alpha_1^{(t_{n-1} - t_{n-p})/\tau} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_p^{(t_n - t_{n-p})/\tau} & \alpha_p^{(t_{n-1} - t_{n-p})/\tau} & \cdots & 1 \end{bmatrix} \begin{bmatrix} r_0^n \\ r_1^n \\ \vdots \\ r_p^n \end{bmatrix} = \mathbf{0}. \tag{4.3}$$

Note that a nontrivial solution must exist, as the matrix is $p \times (p+1)$, and generically the kernel is one-dimensional. The normalisation needs to be established and we quickly find that $r_0^n = 1$ will not work if $p > 1$. For example, take

$$\begin{gathered} \alpha_1 = \mathrm{i}, \quad \alpha_2 = -\mathrm{i} \\ t_n = 3, \quad t_{n-1} = 2, \quad t_{n-2} = 0, \quad \tau = 1 \end{gathered}$$

The matrix equation now reads

$$\begin{bmatrix} -\mathrm{i} & -1 & 1 \\ \mathrm{i} & -1 & 1 \end{bmatrix} \begin{bmatrix} r_0^n \\ r_1^n \\ r_2^n \end{bmatrix} = \mathbf{0}, \text{ implying } r_0^n = 0.$$

Thus requiring $r_0^n = 1$ would make the r_1^n and r_2^n infinite. The situation arises whenever we try to predict a sinusoid from observations taken at or near its zero-crossings; the above example occurs when the sinusoid is given by $\sin \frac{1}{2}\pi t$. We shall return to this point in Chapter 5 when we address the subject of nonlinear prediction, but in the mean time we must find a method of normalising the generalised prediction coefficients so that the total prediction error energy—the device that we intend to minimise in order to fit a set of poles to a set of data—depends continuously on the poles and the sampling instants.

The key step is to constrain the *length* of $\mathbf{r}$ by imposing a normalisation such as the following:

$$\sum_{j=0}^{p} |r_j^n|^2 = N(\boldsymbol{\alpha}). \tag{4.4}$$

Recalling that for regular sampling $N(\boldsymbol{\alpha})$ is just the sum of the squared moduli of the AR coefficients, we see that in that case the definitions (4.2,4.3,4.4) produce the same prediction errors as the classical definition (4.1), up to a factor of modulus 1. This does not matter because we will only need the generalised prediction error *energies* $|\mathfrak{f}_n|^2$.

At this point we may remark that, for given data (x_n, t_n), we have a method for determining the frequencies of a set of p pure tones, simply by minimising $\sum_n |\mathsf{f}_n(\mathbf{x};\boldsymbol{\alpha})|^2$ w.r.t. $\boldsymbol{\alpha}$. The minimum of this function is 0, achieved only when the α_i are correct. Further there does not seem to be any particular reason why the definitions above cannot be used for any AR process, enabling a general AR process to be fitted. This intuition turns out to be well-founded, as we shall see.

We conclude this subsection by discussing why the above construction gives a prediction error energy that is continuous in the model poles and in the sampling instants. The matrix equation (4.3) gives $\mathbf{r} = (r_j)_{j=0}^p$ up to a scaling factor, except in non-generic cases when the matrix has a kernel of dimension > 1. Hence $\mathbf{r}$ is an element of the complex projective space $\mathbb{G}_p$ defined as the quotient space

$$\mathbb{G}_p = \mathbb{C}^{p+1}/\sim, \qquad \mathbf{v} \sim \mathbf{w} \Leftrightarrow (\mathbf{v} = \lambda\mathbf{w},\ 0 \neq \lambda \in \mathbb{C}).$$

$\mathbb{G}_p$ is a compact analytic manifold. By use of elementary row and column operations to reduce the matrix (4.3) to diagonal form, we see that $\mathbf{r} \in \mathbb{G}_p$ is a rational function of the matrix elements, and hence is continuous (viewed as a function into $\mathbb{G}_p$). The following mapping, in which (X_j) are just symbols, is well-defined, because it is invariant under a scaling of $\mathbf{r}$, and it is continuous:

$$\mathbf{r} \in \mathbb{G}_p \mapsto \left((X_j)_{j=0}^p \mapsto \frac{\left|\sum_{j=0}^p r_j X_j\right|^2}{\sum_{j=0}^p |r_j|^2} \right)$$

Also $N(\boldsymbol{\alpha})$ is a continuous function of $\boldsymbol{\alpha}$. The RHS is, up to a factor of $N(\boldsymbol{\alpha})$, the generalised prediction error energy—and that completes the proof. Essentially the space of normalised predictors is compact if we choose the normalisation (4.4), but not if we choose to require $r_0 = 1$.

4.1.3 Prediction errors for white input

Let us construct the generalised prediction errors for a white-noise input $\mathbf{w}$ (variance σ_w^2) and AR model $\boldsymbol{\alpha}$:

$$\mathsf{f}_n(\mathbf{w};\boldsymbol{\alpha}) = \sum_{j=0}^{p} r_j^n w_{n-j}.$$

Apart from the obvious fact that they form an MA(p) process these have the property

$$\mathcal{E}^w |\mathsf{f}_n(\mathbf{w};\boldsymbol{\alpha})|^2 = \left(\sum_{j=0}^{p} |r_j^n|^2 \right) \sigma_w^2 = N(\boldsymbol{\alpha})\sigma_w^2$$

independently of n and hence of the sampling; this implies the weaker result

$$\mathcal{E}_n |\mathsf{f}_n(\mathbf{w};\boldsymbol{\alpha})|^2 = N(\boldsymbol{\alpha})\sigma_w^2,$$

that is, the ratio of prediction error to white input power[1] is $N(\boldsymbol{\alpha})$, just as it is in the classical case[2].

[1] Time-average of the modulus-squared.

[2] The second equation is weaker because it involves a time-average over n. Both equations involve averaging over $\mathbf{w}$; as $\mathbf{w}$ is ergodic, it is immaterial that the first equation uses a realisation-average and the second a time-average over it.

4.1.4 Why 'generalised' prediction?

The previous discussion has seen a departure from the classical theory of linear prediction, for two reasons. First, as $r_0^n \neq 1$ in general (it might even be 0, as in the above example), we cannot view the quantity $-\sum_{j=1}^{p} r_j^n y_{n-j}$ as a prediction for the element y_n. Secondly, only the modulus of $\mathfrak{f}_n$ is defined. Consequently it is better to think of $\mathfrak{f}_n$ as a measure of how well the sequence of $p+1$ data points $(y_{n-p}, \ldots, y_n)$ accords with the pth order AR model $\boldsymbol{\alpha}$. Of course, the smaller $|\mathfrak{f}_n|$ is, the closer the fit.

4.1.5 The incoherent case

Moving to the incoherent case, we now show that the definition (4.3) is the correct one whether or not the poles are on the unit circle. To do this we consider the continuous-time AR process (Ito diffusion[3])

$$\{\mathcal{L}y(t)\}dt = dB(t), \qquad \mathcal{L} \equiv \sum_{j=0}^{p} b_j \left(\frac{d}{dt}\right)^{p-j} \tag{4.5}$$

in which $B(t)$ is a Brownian motion on $\mathbb{R}$ or $\mathbb{C}$. The solution is

$$y(t) = \int_{-\infty}^{\infty} G(t-t')\, dB(t') \tag{4.6}$$

where $G(t)$ is the Green's function satisfying

$$\mathcal{L}G(t) = \delta(t), \qquad t < 0 \Rightarrow G(t) = 0. \tag{4.7}$$

For a stable process, $G(t) \to 0$ exponentially as $t \to \infty$. For an unstable process the integral does not exist. The model (4.5) is an all-pole model in that the Laplace transform of y is a polynomial in the transform-variable ς; the poles are, by elementary arguments, the roots β_i of the polynomial equation $\sum_{j=0}^{p} b_j \varsigma^{p-j} = 0$. The Green's function is given by

$$G(t) = \begin{cases} 0, & t < 0 \\ \displaystyle\sum_{i=1}^{p} \lambda_i e^{\beta_i t}, & t > 0 \end{cases}$$

for appropriate coefficients λ_i which we need not find explicitly.

Now let us sample $y(t)$, so $y_n = y(t_n)$. Then

$$\mathfrak{f}_n(\mathbf{y};\boldsymbol{\beta}) = \sum_{j=0}^{p} r_j^n \int_{-\infty}^{\infty} G(t_{n-j} - t')\, dB(t')$$

[3]Ito diffusions are used for modelling a wide variety of phenomena, including biological systems, mechanics, fluid dynamics, and the pricing of various financial instruments such as 'derivatives'. See [157, 110], both of which go into the subject in much more depth than is required here.

which after expanding the Green's function may be written

$$\mathfrak{f}_n(\mathbf{y};\boldsymbol{\beta}) = \sum_{j=0}^{p}\left(\int_{t_{n-p}}^{t_{n-j}} + \int_{-\infty}^{t_{n-p}}\right)\sum_{i=1}^{p}\lambda_i e^{\beta_i(t_{n-j}-t')}dB(t')r_j^n.$$

By requiring

$$\sum_{j=0}^{p} e^{\beta_i t_{n-j}} r_j^n = 0 \qquad (1 \le i \le p) \tag{4.8}$$

the terms arising from the second integral, i.e. that going back to $t' = -\infty$, vanish. Consequently

$$\mathfrak{f}_n(\mathbf{y};\boldsymbol{\beta}) = \int_{t_{n-p}}^{t_n} (\text{Fn. of } t')\, dB(t').$$

Using the overlap formula for stochastic integrals,

$$\mathcal{E}^B\left\{\int f(t)dB(t)\right\}^* \int g(t)dB(t) = \sigma_B^2 \int f(t)^* g(t)\, dt,$$

we can see that if $f(t)$ and $g(t)$ have non-overlapping support then the LHS must vanish. Applying this to the sequence $\{\mathfrak{f}_n\}$, we have that

- $\{\mathfrak{f}_n\}$ is a MA$(p-1)$ process: $m \ge p \Rightarrow \mathcal{E}^B \mathfrak{f}_n^* \mathfrak{f}_{n+m} = 0$;
- $\{\mathfrak{f}_n\}$ is independent of past observations: $m \ge p \Rightarrow \mathcal{E}^B \mathfrak{f}_n^* y_{n-m} = 0$.

It is a standard result [11] that regular sampling of $y(t)$ produces an ARMA$(p, p-1)$ process—the proof can be effected as above in fact—which implies that after pth order linear prediction the residuals are a MA$(p-1)$ process. The definitions (4.2,4.3) therefore appear to be a natural generalisation of prediction to irregular sampling in the incoherent case, because (4.3) and (4.8) are equivalent if $\alpha_i = e^{\beta_i \tau}$.

4.1.6 Backward prediction errors

If an AR process is stable then its time-reverse is also stable with complex-conjugate AR coefficients. We can then deduce from (4.3) the following expression for the generalised backward prediction errors:

$$\mathfrak{b}_n(\mathbf{y};\boldsymbol{\alpha}) = \sum_{j=0}^{p}(s_j^n)^* y_{n+j}$$

$$\begin{bmatrix} \alpha_1^{(t_{n+p}-t_n)/\tau} & \alpha_1^{(t_{n+p}-t_{n+1})/\tau} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_p^{(t_{n+p}-t_n)/\tau} & \alpha_p^{(t_{n+p}-t_{n+1})/\tau} & \cdots & 1 \end{bmatrix} \begin{bmatrix} s_0^n \\ s_1^n \\ \vdots \\ s_p^n \end{bmatrix} = \mathbf{0} \tag{4.9}$$

and the same normalisation as (4.4) is used to fix the size of $\mathbf{s}^n$. As in §4.1.2 we prefer to think of these as a measure of how well the sequence of data points $y_n, \ldots, y_{n-p}$ accords with the AR model $\boldsymbol{\alpha}^*$. One might ask whether there is a difference between

the forward and backward prediction errors, given that the coefficients r_0^n and s_0^n are not 1. There is a difference, and it occurs because the 'forward' and 'backward' models are both fitted to the data: equivalently, the model is fitted to the data and the conjugate AR model (obtained by conjugating the poles) is fitted to the time-reversed data.

4.1.7 Choice of τ

The value of τ, while not being critical, needs to be chosen with reasonable care so as to avoid problems with multivalued functions.

When we perform the calculations needed to construct the matrices (4.3,4.9), we either work in Cartesian coordinates, so that the real and imaginary parts of α are held, or we work in polars. In the first case the power function $\alpha \mapsto \alpha^t$ is ill-defined unless we cut the plane, and the most natural place to put the branch cut is along $\mathbf{R}^-$. In the second, we need for practical reasons to constrain the argument to prevent the α_i from wandering over the Riemann surface of the log function (which, looking as it does like a spiral staircase, is not bounded). Then a range $(-\pi, \pi)$ is appropriate. Recalling now that $z = e^{2\pi i f \tau}$ we have

$$\arg z \in (-\pi, \pi) \iff f \in (-1/2\tau, 1/2\tau)$$

and so $\mathbf{R}^-$ in the z-plane corresponds to the hypothetical Nyquist frequency $\pm 1/2\tau$.

It is worth mentioning that in the missing data problem, where τ is the lattice interval, the exponents in the matrices of (4.3,4.9) are integers—in which case we can (and should) work in the z-domain and the question of branch cuts does not arise.

4.2 Fitting the model

4.2.1 Method

Now suppose that we have an irregularly-sampled dataset with x_n the observations and t_n the times at which they are observed (we assume $m < n \Rightarrow t_m < t_n$). The first thing to do is to set τ. Then, for a set of poles $\boldsymbol{\alpha}$ we construct the forward and backward prediction coefficients r_j^n, s_j^n and prediction errors $\mathfrak{f}_n, \mathfrak{b}_n$ and thence the total forward and backward prediction error power

$$E(\mathbf{x}; \boldsymbol{\alpha}) \stackrel{\text{def}}{=} \frac{1}{2(N-p)} \sum_{n=p+1}^{N} |\mathfrak{f}_n(\mathbf{x}; \boldsymbol{\alpha})|^2 + |\mathfrak{b}_{n-p}(\mathbf{x}; \boldsymbol{\alpha})|^2. \tag{4.10}$$

We then minimise $E(\mathbf{x}; \boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$. As discussed in §4.1.7, we constrain the arguments to lie within $(-\pi, \pi)$ (or equivalently stop the α_i wandering across the branch cut). It is not necessary to constrain the moduli of the α_i, because the inclusion of backward prediction error energies in the cost function keeps the model stable. The way that we have performed the minimisation is this. Our computer program has been developed for modelling real-valued data. We assume an even

model order and write the poles in complex conjugate pairs: $\alpha_1 = \alpha_2^*$, $\alpha_3 = \alpha_4^*$, etc. First the performance surface is searched, with

$$0 < \arg\alpha_1 < \arg\alpha_3 < \cdots < \pi$$
$$0.95 < |\alpha_i| < 1.00$$

The moduli are chosen randomly between 0.95 and 1.00; the arguments are stepped round with a granularity of approximately $\pi/6$ radians. The 'best' point, i.e. that with the minimum E-value, is selected as the starting-point for a simple gradient-descent algorithm. We estimate the local gradient at a certain point by looking at nearby points, and then walk down the performance surface in small steps in the direction of steepest descent. When no further progress can be made, the step size is made smaller and the process is repeated. The resulting minimum is found to a precision $|\delta\alpha| < 0.001$. It might appear that, by choosing all the starting points very near the unit circle, one cannot obtain a model with poles away from the unit circle; but this is not so. Unnecessary poles are moved away from the unit circle during the optimisation. Hence a 6th-order model may have a spectrum with three, two or one pronounced spikes, or even none at all. This is seen in the test results.

4.2.2 A note on the minimum error energy

Let us investigate the relation between generalised forward and backward prediction error energies for the model orders p and $p+1$. For model order p the prediction coefficients are given by (4.2). Let us now increase the model order by 1, introducing a $(p+1)$th pole α_{p+1} at the origin. Call the new prediction coefficients $\overline{r}_j^n$ and the new prediction errors $\overline{\mathfrak{f}}_n$ and $\overline{\mathfrak{b}}_n$. By (4.3),

$$\begin{bmatrix} \alpha_1^{(t_n - t_{n-p-1})/\tau} & \alpha_1^{(t_{n-1} - t_{n-p-1})/\tau} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_p^{(t_n - t_{n-p-1})/\tau} & \alpha_p^{(t_{n-1} - t_{n-p-1})/\tau} & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \overline{r}_0^n \\ \overline{r}_1^n \\ \vdots \\ \overline{r}_{p+1}^n \end{bmatrix} = \mathbf{0}.$$

Clearly $\overline{r}_{p+1}^n = 0$ from this; now for each i multiply the ith row by $\alpha_i^{(t_{n-p} - t_{n-p-1})/\tau}$ to get

$$\begin{bmatrix} \alpha_1^{(t_n - t_{n-p})/\tau} & \alpha_1^{(t_{n-1} - t_{n-p-1})/\tau} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_p^{(t_n - t_{n-p})/\tau} & \alpha_p^{(t_{n-1} - t_{n-p-1})/\tau} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \overline{r}_0^n \\ \overline{r}_1^n \\ \vdots \\ \overline{r}_p^n \end{bmatrix} = \mathbf{0}.$$

Noting the obvious fact

$$N(\alpha_1, \ldots, \alpha_p, 0) \equiv N(\alpha_1, \ldots, \alpha_p)$$

we have $|\overline{\mathfrak{f}}_n|^2 = |\mathfrak{f}_n|^2$. By similar arguments $|\overline{\mathfrak{b}}_n|^2 = |\mathfrak{b}_n|^2$. We may conclude that the introduction of a new pole at the origin does not affect the prediction error energies.

Therefore the minimum prediction error energy for model $p+1$ cannot be any higher than that for model order p. Writing E_p for the minimum assumed by $E(\mathbf{x};\boldsymbol{\alpha})$ as $\boldsymbol{\alpha} \in \mathbb{C}^p$ is varied, we can write this statement as

$$E_0 \geq E_1 \geq E_2 \geq \cdots$$

It is not difficult to see that $E_0 = N^{-1}\sum_{n=1}^{N} |x_n|^2$. Also, we have a generalisation of the various order selection criteria mentioned in (2.6), for $\hat{\sigma}_\varepsilon^2[p]$, the estimate of the driving noise variance for model order p, is $E_p(\mathbf{x})$.

4.2.3 Properties of $E(\mathbf{x};\boldsymbol{\alpha})$

This section summarises the main properties of E.

- In the coherent case E is zero precisely when the α_i are in the correct places ($\alpha_i = e^{i\omega_i\tau}$); so of course minimising E gives the right answer *even for finite data length.*
- When the sampling is regular $\tau = \delta t$, and so our method reduces to an unconstrained minimisation of forward and backward (classical) prediction error energy—the Covariance method.
- Given that the solution is unique[4] in the case of regular sampling, and that the generalised prediction coefficients depend continuously on the sampling instants, we can expect $\boldsymbol{\alpha} \mapsto E(\mathbf{x},\boldsymbol{\alpha})$ to be free from local minima when the samples deviate 'slightly' from uniformity [5]. Given that the function $\boldsymbol{\alpha} \mapsto E(\mathbf{x},\boldsymbol{\alpha})$ is infinitely differentiable as well, there must be reasonable optimism that minimising it will not be too onerous.
- With the notation of §4.2.2,

$$E_0 \geq E_1 \geq E_2 \geq \cdots$$

so that E_p/E_0 provides a measure of goodness-of-fit.

4.3 A description of the AR(1) case

At this point we have completed a description of a method for representing a time series by a set of poles (α_i) evaluated for a specific Nyquist limit $1/2\tau$. Before moving

[4] We mean unique up to a permutation of the poles.

[5] As each prediction error is constructed from $p+1$ data points, we mean that each $(p+1)$-tuple of points should not deviate too far from uniformity. This is therefore a 'local' definition of uniformity. The distinction between this and a global constraint on sampling uniformity is the same as the distinction between additive-random and jittered sampling. One may additive-randomly sample a time series using an intersample pdf that has narrow width, i.e. 'almost a delta-function', and this will give rise to a sequence of observations that is locally quite uniform. But over a long stretch of data it will be quite nonuniform. Our comment on the function E is that it should have a unique minimum when the samping deviates slightly from uniformity in either the additive-random or the jittered cases.

on to consider some test examples, let us consider the first-order continuous-time stochastic process (Ornstein-Uhlenbeck, or OU, process)

$$dy(t) + b'y(t)dt = dB(t), \qquad \mathrm{Re}\, b' > 0 \tag{4.11}$$

which is solved by the stochastic integral

$$y(t) = \int_{-\infty}^{t} e^{b'(t'-t)} dB(t').$$

Following the nomenclature that we have established,

$$y_n = e^{-b'(t_n - t_{n-1})} y_{n-1} + e^{-b'(t_n - t_{n-1})} \int_{t_{n-1}}^{t_n} e^{b'(t' - t_{n-1})} dB(t').$$

First let us consider the case when we know b' and merely wish to simulate observations from this process. For brevity let us write f_n for the last term in the above equation. Of course it has a simple interpretation as the forward innovation error and, from the discussions in §4.1.6, it satisfies the equations

$$m \geq 1 \Rightarrow \mathcal{E}^B f_n^* y_{n-m} = \mathcal{E}^B f_n^* f_{n \pm m} = 0$$

We can go further, for

$$\begin{aligned} \mathcal{E}^B |f_n|^2 &= \frac{\sigma_B^2}{2\mathrm{Re}\, b'} \left(1 - e^{2\mathrm{Re}\, b' . (t_{n-1} - t_n)}\right) \\ \mathcal{E}^B |y_n|^2 &= \frac{\sigma_B^2}{2\mathrm{Re}\, b'} \end{aligned}$$

and so to simulate a first-order stochastic process of variance σ_y^2 and decay parameter b' we have only to calculate

$$y_n = e^{-b'(t_n - t_{n-1})} y_{n-1} + \sigma_y \left(1 - e^{2\mathrm{Re}\, b' . (t_{n-1} - t_n)}\right)^{1/2} g_n$$

where the g_n are independent observations from the standard Normal distribution.

Now let us consider how our generalised prediction error approach fares when we wish to estimate b' from a section of data. By (4.3)

$$\begin{bmatrix} r_0^n \\ r_1^n \end{bmatrix}, \begin{bmatrix} s_0^n \\ s_1^n \end{bmatrix} \propto \begin{bmatrix} 1 \\ -e^{b(t_{n-1} - t_n)} \end{bmatrix}$$

Suppose that, rather than the normalisation (4.4), we simply put $r_0^n = s_0^n = 1$ (which is always alright when $p = 1$ but, as we have seen, goes wrong when $p > 1$). Then

$$\begin{aligned} \mathfrak{f}_n(\mathrm{y}, b) &= y_n - e^{b(t_{n-1} - t_n)} y_{n-1} \\ \mathfrak{b}_{n-1}(\mathrm{y}, b) &= y_{n-1} - e^{b(t_{n-1} - t_n)} y_n \end{aligned}$$

and because

$$\mathcal{E}^B y_{n-m}^* y_n = \frac{\sigma_B^2}{2\mathrm{Re}\, b'} e^{b'(t_{n-m} - t_n)}, \qquad m \geq 0$$

their variances are

$$\mathcal{E}^B|\mathrm{f}_n(\mathbf{y},b)|^2 = \mathcal{E}^B|\mathrm{b}_{n-1}(\mathbf{y},b)|^2 = \frac{\sigma_B^2}{2\mathrm{Re}\, b'}\left(1+\left|e^{b(t_{n-1}-t_n)}\right|^2 - 2\mathrm{Re}\left\{e^{(b^*+b')(t_{n-1}-t_n)}\right\}\right).$$

For brevity write $\mathrm{PE}(b)$ for this quantity. Then

$$\mathrm{PE}(b') = \frac{\sigma_B^2}{2\mathrm{Re}\, b'}\left(1-\left|e^{b'(t_{n-1}-t_n)}\right|^2\right)$$

and so

$$\mathrm{PE}(b) - \mathrm{PE}(b') = \frac{\sigma_B^2}{2\mathrm{Re}\, b'}\left(\left|e^{b(t_{n-1}-t_n)} - e^{b'(t_{n-1}-t_n)}\right|^2\right) \geq 0.$$

This is an important conclusion, because we can now say that 'on average' (i.e. over all realisations of the underlying random walk $B(t)$)

$$E(\mathbf{y},b) \geq E(\mathbf{y},b')$$

and it follows that minimising $E(\mathbf{y},b)$ with respect to b is a good way to estimate b'. The same conclusion is obtained even if the prediction error energies are 'weighted' with weights depending on the observation times. One might, for example, notice that $\mathrm{PE}(b')$ is approximately proportional to $(t_n - t_{n-1})$, and minimise the function

$$E(\mathbf{y},b) \stackrel{\text{new def}}{=} \frac{1}{2(N-1)}\sum_{n=2}^{N}\frac{|\mathrm{f}_n(\mathbf{y},b)|^2 + |\mathrm{b}_{n-1}(\mathbf{y},b)|^2}{t_n - t_{n-1}}.$$

It is worth noting that the normalisation (of the predictors) used in the derivation of the optimality result $E(\mathbf{y},b) \geq E(\mathbf{y},b')$ is not the same as that posed for the general case (4.4). That raises the question of whether (4.4) is the right normalisation to use. Unfortunately it does not raise a satisfactory answer, for the problem is almost intractable for $p > 1$. It may be that (4.4) is not quite the best choice, but we are going to show by way of examples in the next section that it works, and that is an important recommendation.

4.4 Test examples

It is almost universal to test autoregressive algorithms on data consisting of sinusoids in noise; by so doing one can investigate the resolution, performance for different SNR, and so on. We have therefore run some tests on this kind of data. One objection to this kind of test is that tones in noise do not obey an AR model. Therefore we have also tested the algorithm in question on genuinely broad-band data.

The specific aim in carrying out these tests has been to examine if, and how, the model fitting procedure depends on the sampling scheme. Ideally there should be no dependence except when there are spectral components above the Nyquist limit; in those cases regular sampling would give rise to aliasing, so the results for different sampling schemes would not agree.

4.4.1 Test on one sinusoid in noise

The first test was to take 64 samples of a sinusoid in white additive Gaussian noise of standard deviation 1. The frequency of the sinusoid was taken (arbitrarily) as 0.27Hz. The quality of fit was assessed as a function of the following three parameters:

- Signal-to-noise ratio. Four values of the amplitude (A) of the sinusoid were considered: 8,4,2 (SNR +15,+9,+3 dB) and 0. The purpose of trying $A = 0$ was to examine the level of spurious features thrown up.
- Model order. Orders 2 and 6 were used.
- Sampling scheme. Six sampling schemes were applied:
  - (a) regular sampling with period 1 (i.e. 1 second)
  - (b) additive random sampling, intersample spacings drawn from the rectangular ('uniform') distribution on $[0.5, 1.5]$
  - (c) additive random sampling, intersample spacings drawn from the $\Gamma(2,2)$ distribution
  - (d) additive random sampling, intersample spacings drawn from the $\Gamma(3,3)$ distribution
  - (e) jittering the sampling in (a), varying the positions of the samples randomly, with uniform probability, up to ± 0.25
  - (f) jittering the sampling in (a), varying the positions of the samples randomly, with uniform probability, up to ± 0.5.

The $\Gamma(\nu, \lambda)$ distribution has pdf

$$\frac{1}{\Gamma(\nu)} \lambda^{\nu} t^{\nu-1} e^{-\lambda t} \qquad (\nu > 0, \lambda > 0)$$

and has mean ν/λ. (When $\nu = 1$ it is the Poisson distribution.) In each of the six cases enunciated above, the average sampling rate is 1Hz. The parameter τ was set to 1 for each of these tests, corresponding to a notional Nyquist interval $[-\frac{1}{2}, \frac{1}{2}]$.

The results are shown in Figures 4.1–4. Some general trends are clear. For SNR=+15dB (corresponding to $A = 8$) the results for model order 2 are all good (and virtually identical); the peak becomes lower and wider as the SNR reduces. The effect of increasing the model order is to sharpen the peak, as is a well-known effect with the classical AR spectral estimator; but the results for sampling schemes (c) and (d) are consistently not as good as for the other four schemes. We are not sure why this should be, but it may be due to the wider spread of intersample spacings for those two schemes.

4.4.2 Test on two sinusoids in noise

The second test was to take 64 samples of a pair of sinusoids, each of amplitude 4, in white additive Gaussian noise of standard deviation 1. The same sampling schemes as in §4.4.1 were applied, and model orders of 4 and 6 were used. See Figures 4.5–6.

First the frequencies were selected as 0.27 and 0.37; one sees that the AR(4) model does not separate the two components, but the addition of a third pair of poles is sufficient to split them.

When the frequencies were altered to 0.43 and 0.83 the uniform sampling gives rise to aliasing, but the nonuniform sampling schemes correctly identify the frequencies (the 'Nyquist interval' was expanded by putting $\tau = 0.5$ for this example). In fact the uniform sampling gave us some problems, because the error function $E(\mathbf{x};\boldsymbol{\alpha})$ had multiple minima and depending on the starting-conditions different answers were produced. Of course, this is not a fault of the method, which if the sampling is regular cannot be expected to differentiate between frequencies above the Nyquist limit from their aliases.

4.4.3 Test on broad-band signals

A 64-point test sequence was created by calculating

$$x(t) = y_c(t)\cos 2\pi f_0 t + y_s(t)\sin 2\pi f_0 t \tag{4.12}$$

with $y_c(t)$ and $y_s(t)$ independent observations from the OU process (4.11) with parameters $b = 0.02$ or 0.20, $\sigma_y^2 = 50$ (see §4.3). Different values of b give different degrees of incoherence, or equivalently different notional bandwidths. The 'centre frequency' f_0 was taken as 0.21. The model order was taken as 2 and the fit was assessed as a function of the sampling scheme. The same six schemes as in §4.4.1 were applied. As can be seen from Figure 4.7, the AR spectra are virtually identical. In particular the results for sampling schemes (c) and (d) are not demonstrably worse, as they are in the results of §4.4.1. Perhaps this is because the signals under consideration in these tests are closer to being autoregressive.

As an alternative method of spectral broadening we have considered amplitude modulation by a continuous-time chaotic signal, the Lorenz attractor [212]:

$$x(t) = Z_{1c}(\gamma t)\cos 2\pi f_0 t + Z_{1s}(\gamma t)\sin 2\pi f_0 t. \tag{4.13}$$

The Lorenz equations are

$$\begin{aligned} dZ_1/dt &= 10(Z_2 - Z_1) \\ dZ_2/dt &= -Z_1 Z_3 + 28Z_1 - Z_2 \\ dZ_3/dt &= Z_1 Z_2 - \tfrac{8}{3}Z_3 . \end{aligned}$$

Z_{1c} and Z_{1s} are 'independent' realisations of the chaotic process, in that although they are both observations on the Z_1 variable their starting conditions are different.

The integration of the Lorenz equations was performed using an adaptive step-size fourth-order Runge-Kutta algorithm[6] and a relative accuracy of 10^{-6}. The

[6] NAG routine D02BAF.

sampling could therefore be performed by nominating the next sample point, and integrating in time as far as that point. The initial conditions, which correspond to a points very close to the attractor (as opposed to arbitrary points in space) were $(Z_{ic})(0) = (-15, -21, 30)$ and $(Z_{is})(0) = (7, 15, 10)$.

A typical realisation of data from the Z_1 variable, and its characteristically broad, continuous spectrum, are shown in Figure 5.4–5. By altering γ we can control the coherence of x : the lower γ is, the more slowly the modulator varies, and the more coherent the process. We have considered two values, 0.02 and 0.10. Again (see Figure 4.8) the AR spectra are seen to be virtually identical.

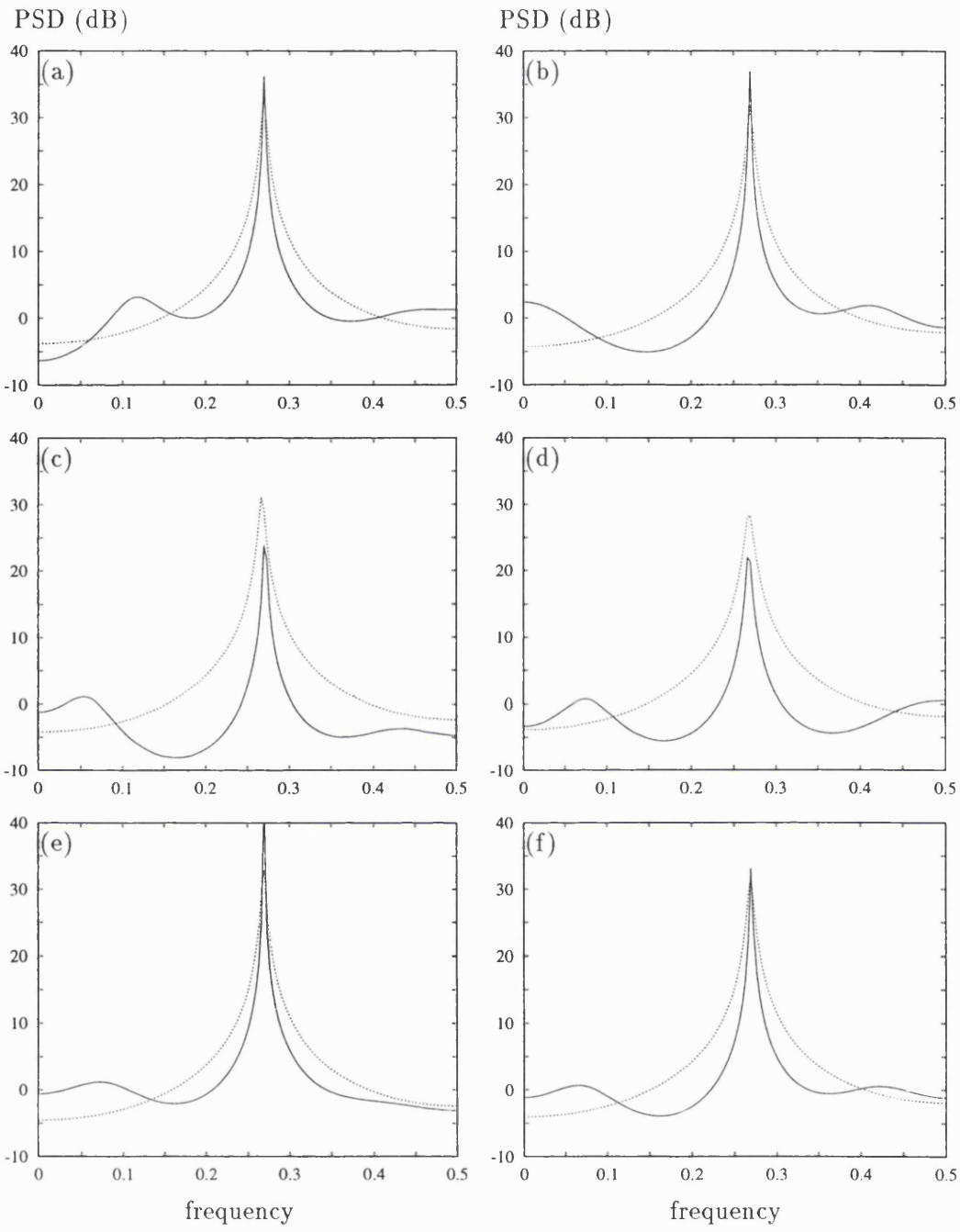


Figure 4.1. AR(p) spectrum for single sinusoid (frequency 0.27) in white noise (SNR +15dB).
Sampling schemes: see §4.4.1. Orders: [solid line] $p = 6$, [dotted line] $p = 2$.

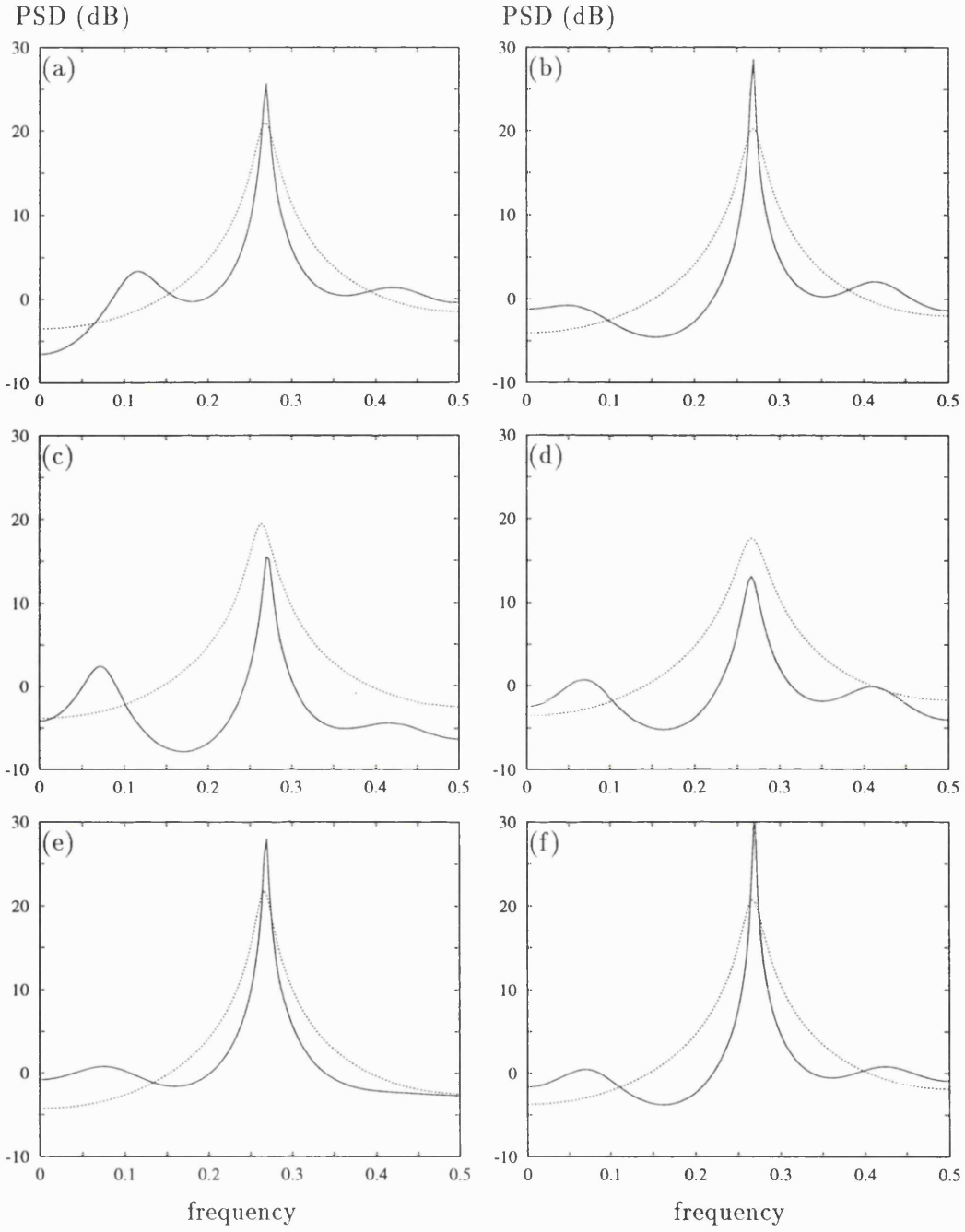


Figure 4.2. AR(p) spectrum for single sinusoid (frequency 0.27) in white noise (SNR +9dB).
Sampling schemes: see §4.4.1. Orders: [solid line] $p = 6$, [dotted line] $p = 2$.

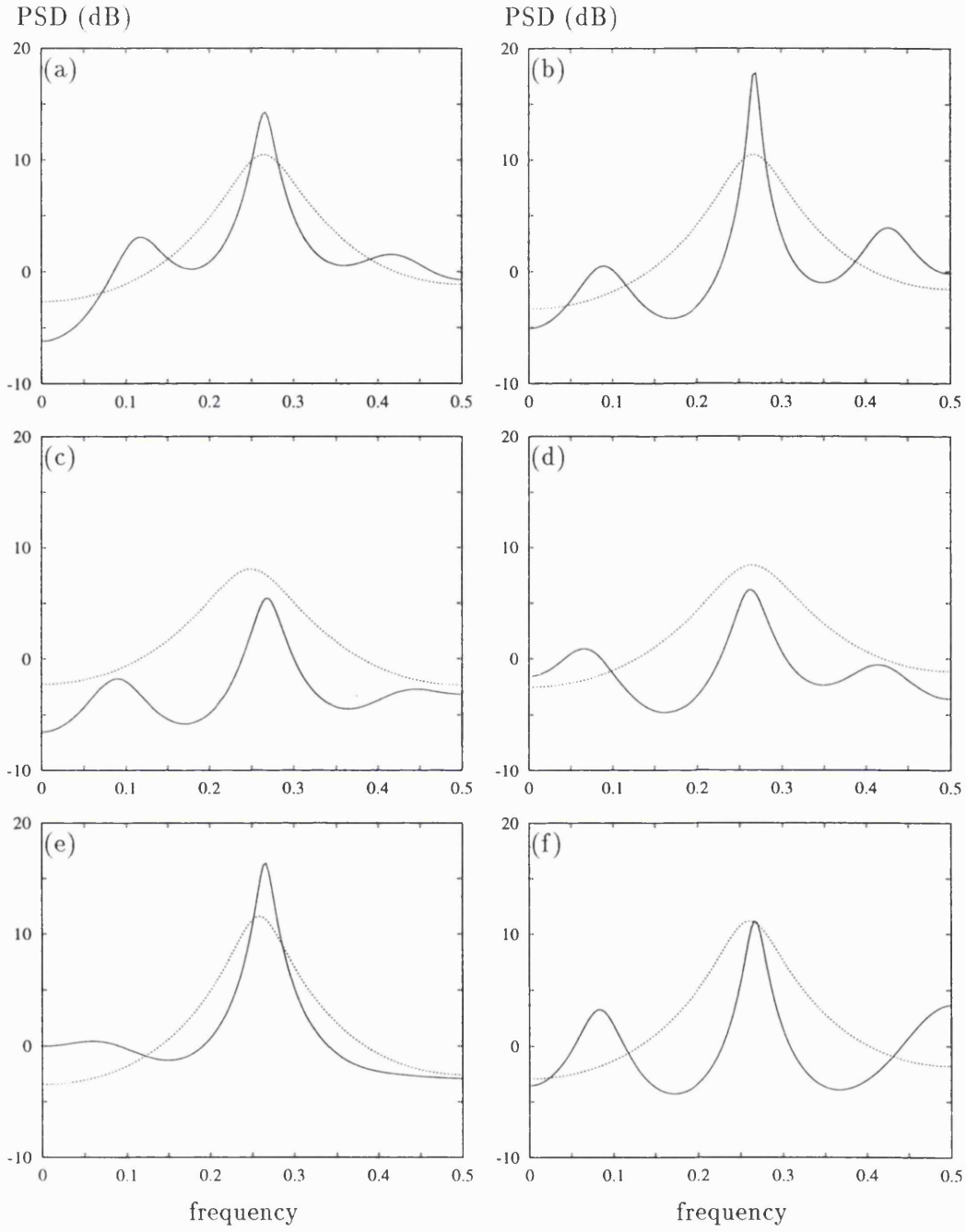


Figure 4.3. AR(p) spectrum for single sinusoid (frequency 0.27) in white noise (SNR +3dB).
Sampling schemes: see §4.4.1. Orders: [solid line] $p = 6$, [dotted line] $p = 2$.

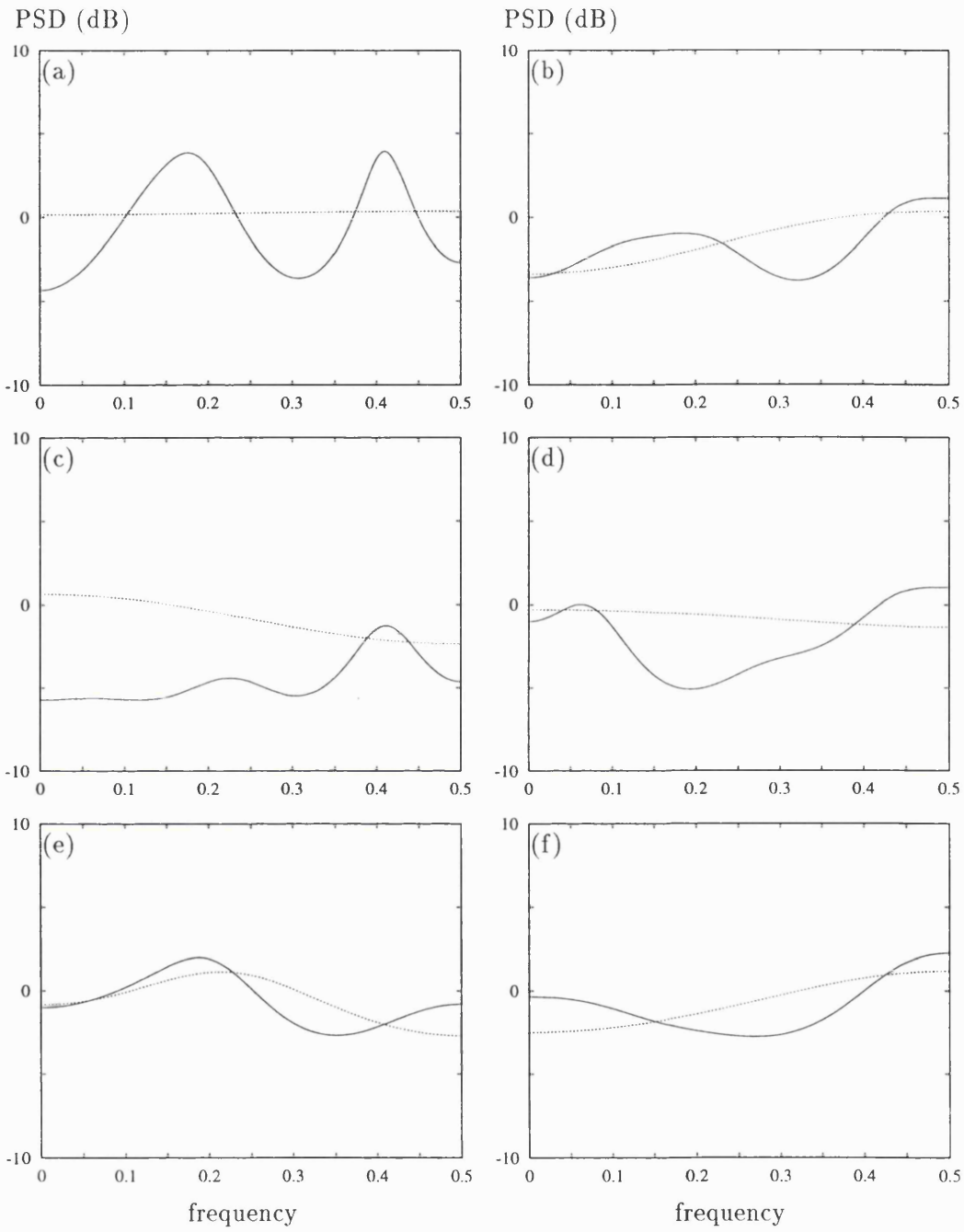


Figure 4.4. AR(p) spectrum for white noise.
Sampling schemes: see §4.4.1. Orders: [solid line] $p = 6$, [dotted line] $p = 2$.

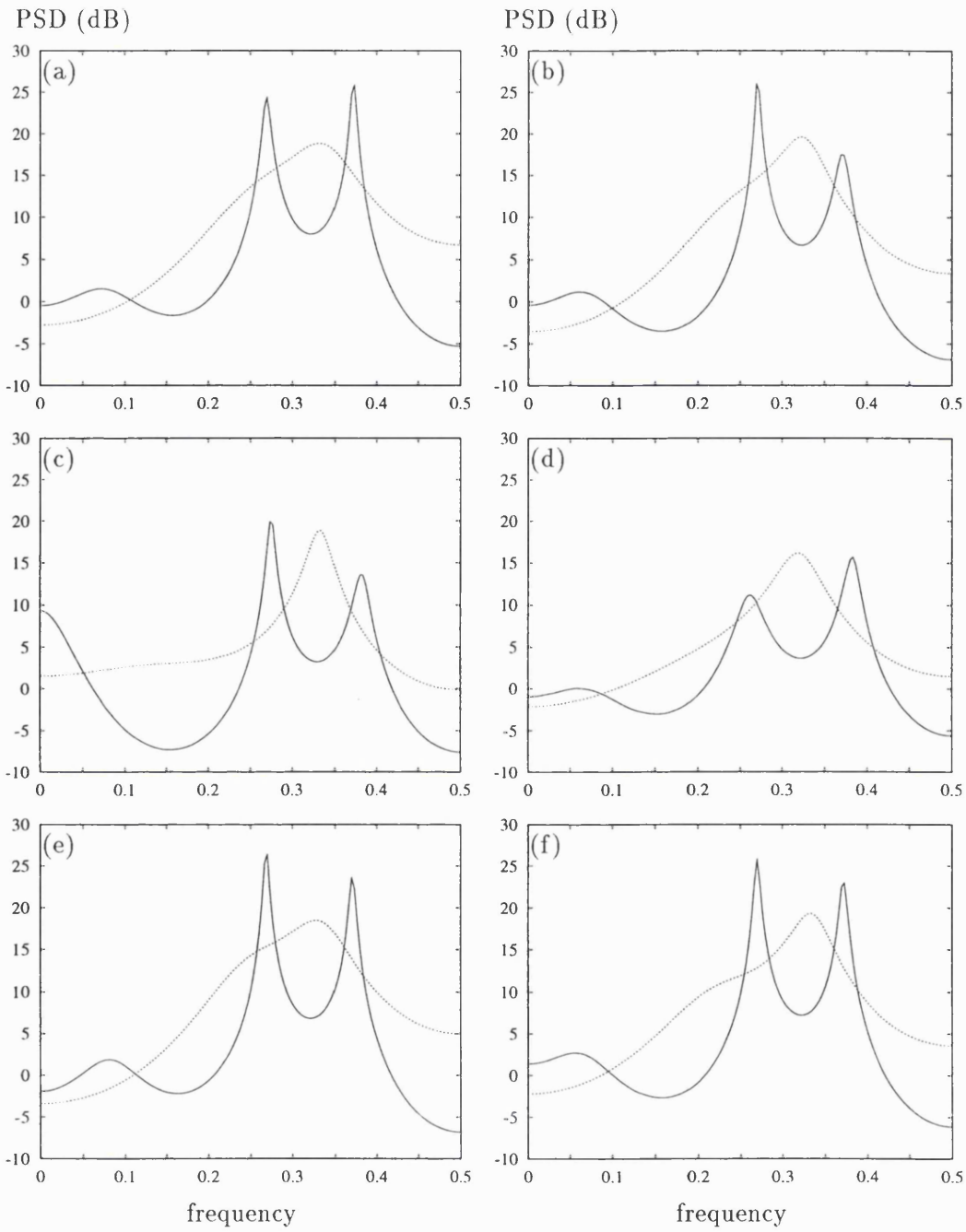


Figure 4.5. AR(p) spectrum for pair of tones in white noise (frequencies 0.27, 0.37; SNR +9dB).
Sampling schemes: see §4.4.1. Orders: [solid line] $p = 6$, [dotted line] $p = 4$.

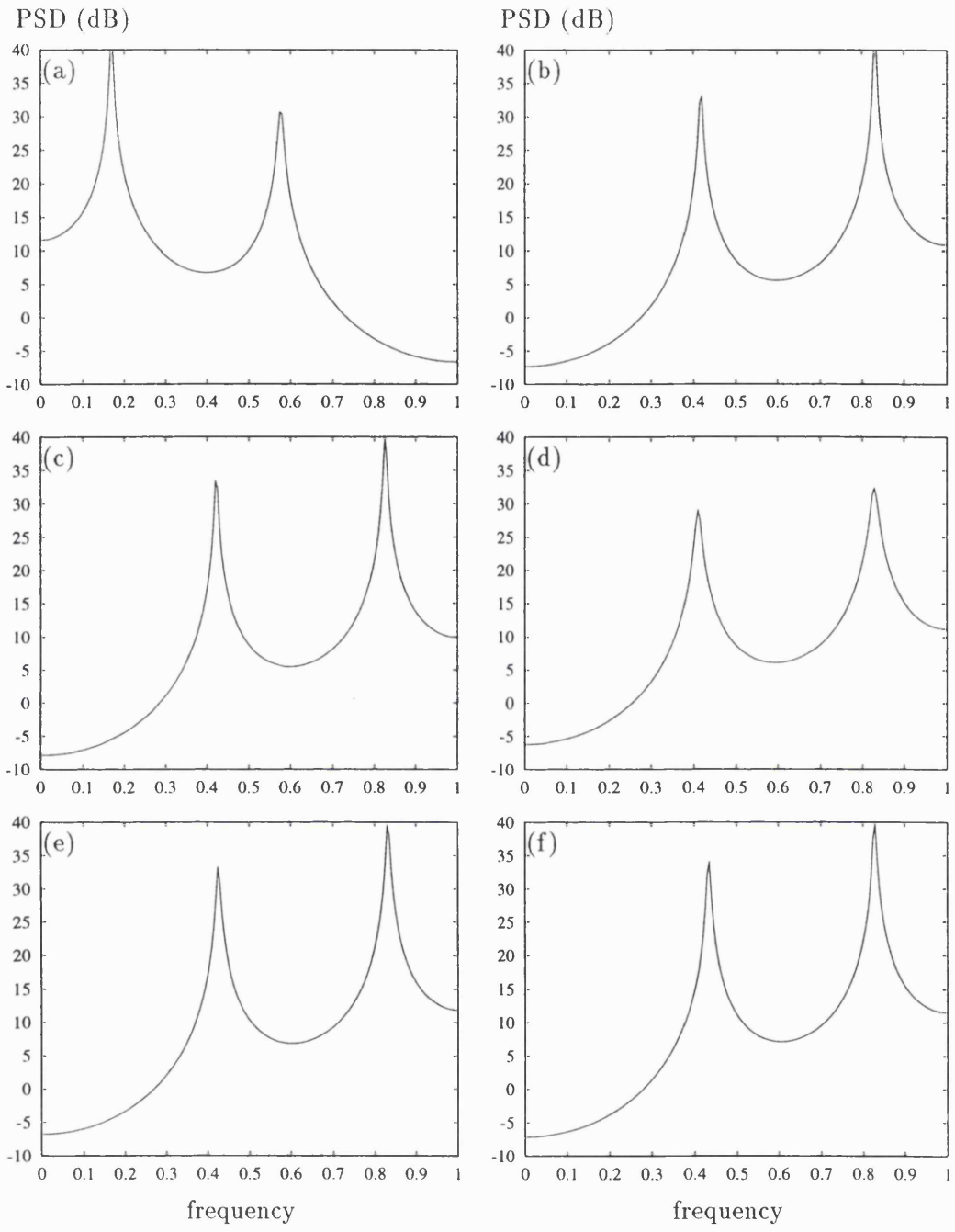


Figure 4.6. AR(p) spectrum for pair of tones in white noise (frequencies 0.43, 0.83; SNR +9dB).
Sampling schemes: see §4.4.1. Orders: [solid line] $p = 6$, [dotted line] $p = 4$.
Note that aliasing has occurred in (a).

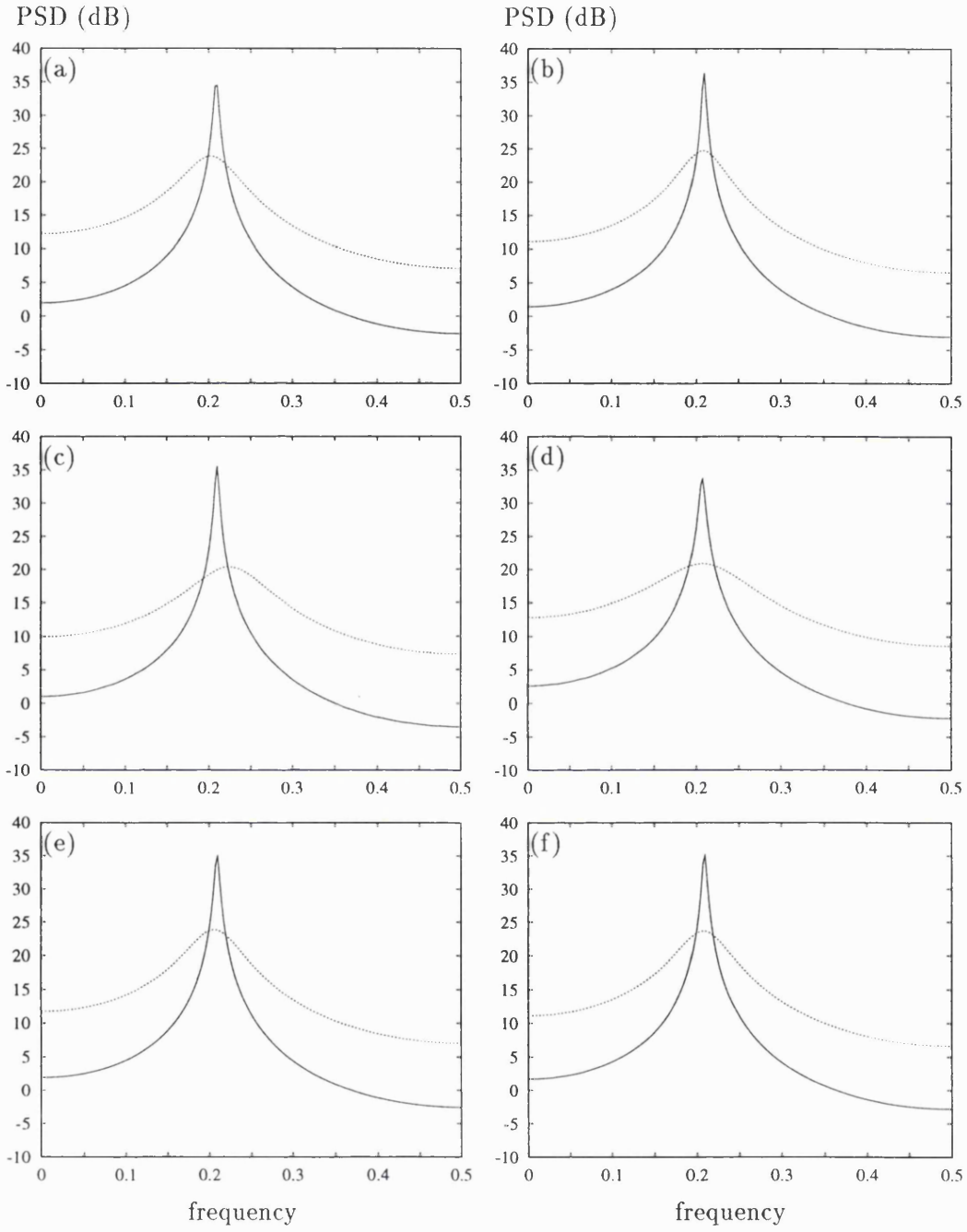


Figure 4.7. AR(2) spectrum for continuous-time stochastic process (4.12). [Solid line] $b = 0.02$, [dotted line] $b = 0.20$. Sampling schemes: see §4.4.1.

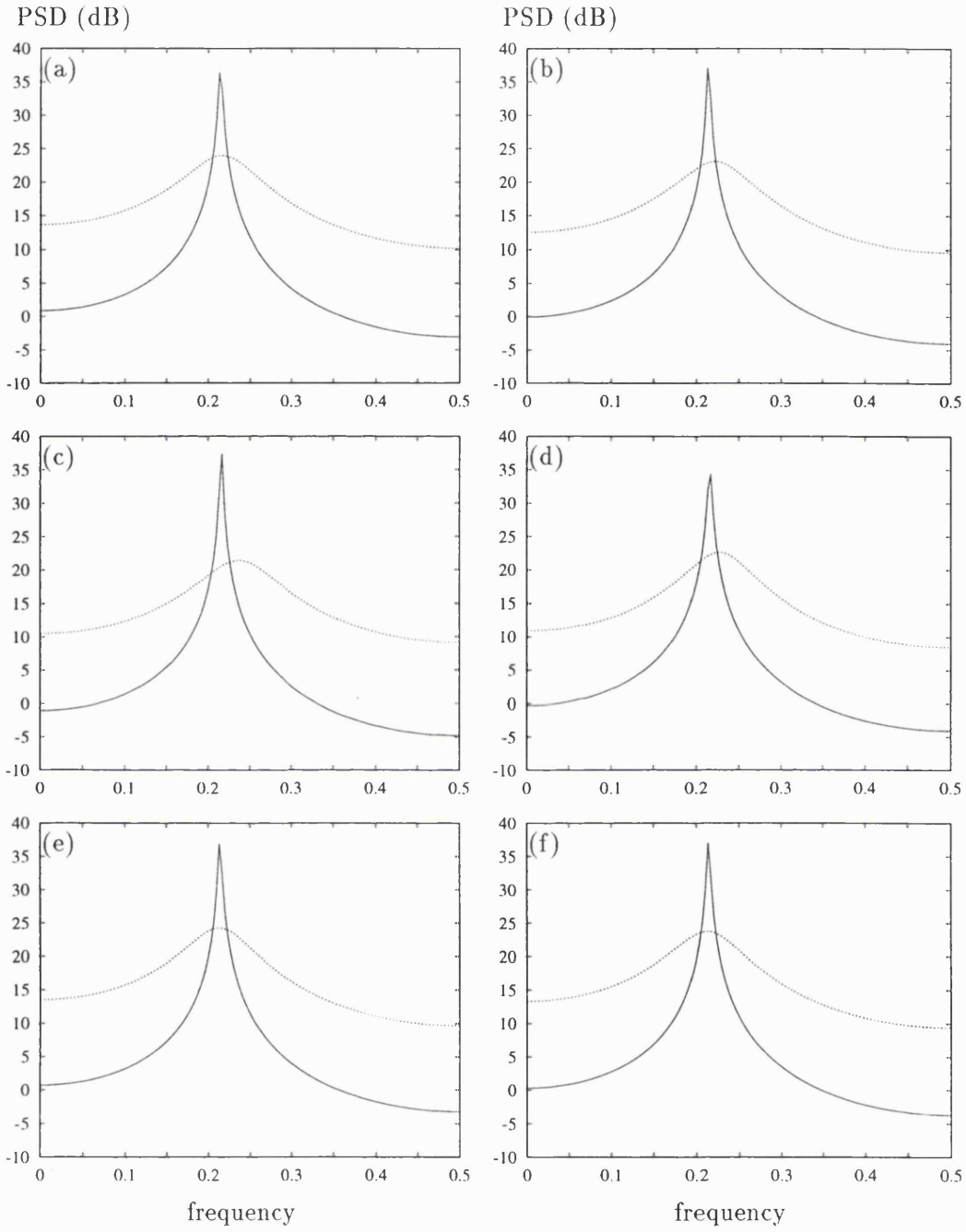


Figure 4.8. AR(2) spectrum for continuous-time dynamical process (4.13). [Solid line] $\gamma = 0.02$, [dotted line] $\gamma = 0.10$. Sampling schemes: see §4.4.1.

4.5 Filtering and separation

4.5.1 The Wiener filter

In this section we shall consider the following problem

> A signal is given consisting of the sum of an AR process with known poles and a 'message signal' about which nothing is known. Do the separation.

followed by its generalisation

> Now do it when the sampling is irregular.

We remark at this point that we are going to characterise an AR process by means of its forward and backward prediction error energy, which (as seen earlier) is generalisable to the irregular case. To start with it is worthwhile to consider the maximum-likelihood separation of two signals $\mathbf{x}$ and $\mathbf{s}$ that have known covariance matrices $\mathbf{C_x}$ and $\mathbf{C_s}$. Their sum is given as $\mathbf{z}$. As before, N is the data length. Assuming the two are independent and Gaussian we maximise

$$(2\pi)^{-N}(\det \mathbf{C_x})^{-1/2}(\det \mathbf{C_s})^{-1/2} \exp -\tfrac{1}{2}\left\{\boldsymbol{\xi}^\dagger \mathbf{C_x}^{-1}\boldsymbol{\xi} + (\mathbf{z}-\boldsymbol{\xi})^\dagger \mathbf{C_s}^{-1}(\mathbf{z}-\boldsymbol{\xi})\right\} \quad (4.14)$$

with respect to $\boldsymbol{\xi}$; this of course can readily be solved to give the Wiener filter (see e.g. [77, 104, 160]), which estimates $\mathbf{x}$ as $\hat{\mathbf{x}}$ by

$$\hat{\mathbf{x}} = \left(\mathbf{I} + \mathbf{C_s}\mathbf{C_x}^{-1}\right)^{-1}\mathbf{z}.$$

Now suppose that $\mathbf{x}$ comes from an AR model whose parameters are known but whose driving noise variance is possibly unknown. Suppose also that nothing is known about $\mathbf{s}$. Then we must assume that $\mathbf{C_s}$ is a multiple of the identity matrix, so $\mathbf{C_x} = \sigma_\varepsilon^2\left(\mathbf{F}^{f\dagger}\mathbf{F}^f\right)^{-1}$ and $\mathbf{C_s} = \sigma_s^2\mathbf{I}$, where $\mathbf{F}^f$ is the forward prediction error matrix,

$$\mathbf{F}^f = \begin{bmatrix} a_p & a_{p-1} & \dots & 1 & & \\ & a_p & a_{p-1} & \dots & 1 & \\ & & \dots & \dots & \dots & \dots \end{bmatrix}$$

(the vacant spaces are supposed to be filled with zeros). Then the solution is

$$\hat{\mathbf{x}} = \left(\mathbf{I} + \frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{F}^{f\dagger}\mathbf{F}^f\right)^{-1}\mathbf{z}.$$

We may wish to use the backward prediction error filter matrix

$$\mathbf{F}^b = \begin{bmatrix} 1 & a_1^* & \dots & a_p^* & & \\ & 1 & a_1^* & \dots & a_p^* & \\ & & \dots & \dots & \dots & \dots \end{bmatrix}$$

as well, and if we write $\mathbf{F} = \begin{pmatrix} \mathbf{F}^f \\ \mathbf{F}^b \end{pmatrix}$ then the solution can be modified to

$$\hat{\mathbf{x}} = \left(\mathbf{I} + \frac{\sigma_s^2}{2\sigma_\varepsilon^2}\mathbf{F}^\dagger\mathbf{F}\right)^{-1}\mathbf{z}. \tag{4.15}$$

(Note that $|\mathbf{F}\mathbf{v}|^2 \equiv 2(N-p)E(\mathbf{v};\mathbf{a})$.) The matrix to be inverted in (4.15) is banded, so the equation can be solved quickly using Cholesky factorisation in $O(Np^2)$ operations. The sticking-point is that we do not know the variance ratio $\sigma_s^2/\sigma_\varepsilon^2$. Let us therefore consider the possibility of introducing a parameter μ for $\sigma_s^2/2\sigma_\varepsilon^2$:

$$\hat{\mathbf{x}}_\mu = \left(\mathbf{I} + \mu\mathbf{F}^\dagger\mathbf{F}\right)^{-1}\mathbf{z}. \tag{4.16}$$

4.5.2 Estimating the variance ratio

We have considered three approaches for finding μ and each involves the introduction of a 'self-consistency' constraint. These are tabulated below.

Name	Constraint
E	$\lvert\mathbf{F}\hat{\mathbf{x}}_\mu\rvert^2 = 2(N-p)\sigma_\varepsilon^2$
C	$\lvert\mathbf{F}\hat{\mathbf{x}}_\mu\rvert^2/\lvert\hat{\mathbf{x}}_\mu\rvert^2 = 2C(\mathbf{a})$
R	$\lvert\mathbf{z}-\hat{\mathbf{x}}_\mu\rvert^2/\lvert\mathbf{F}\hat{\mathbf{x}}_\mu\rvert^2 = \mu$
B	C or R (see text)

These can be justified as follows. In the E case, we know that $\mathbf{x}$ comes from an AR process with known parameters and variance, and hence the total forward and backward prediction error energy has expectation $2(N-p)\sigma_\varepsilon^2$. In the C case this energy is unknown but it is proportional to the signal energy and the proportionality constant is $2C(\mathbf{a})$. In $\mathbb{R}^N$ the surfaces described by the constraints are an ellipsoid and a cone. The R method is rather different: given that μ supposedly estimates $\sigma_s^2/2\sigma_\varepsilon^2$ and that $|\mathbf{z}-\hat{\mathbf{x}}_\mu|^2/|\mathbf{F}\hat{\mathbf{x}}_\mu|^2$ also does, the two had better be equal. In fact we can also obtain this equation by maximising the likelihood (4.14) w.r.t. μ.

What we shall do now is to show that all three types of solution can be obtained by repeated bisection (or a variant thereof[7]), thereby providing a robust method of solution. It is convenient to diagonalise $\mathbf{F}^\dagger\mathbf{F}$, so let us consider[8] the singular-value decomposition (SVD) of $\mathbf{F}$, i.e.

$$\mathbf{F} = \mathbf{Q}\begin{pmatrix}\mathbf{D}\\ \mathbf{O}\end{pmatrix}\mathbf{P}^\dagger, \qquad \mathbf{Q}\mathbf{Q}^\dagger = \mathbf{P}\mathbf{P}^\dagger = \mathbf{I}, \quad \mathbf{D} = \mathrm{diag}(\delta_1^{\frac{1}{2}},\ldots,\delta_N^{\frac{1}{2}})$$

[7]The *ne plus ultra* of bisection algorithms is the van Wijngaarden–Dekker–Brent method, which combines superlinear convergence with the robustness of the simple bisection method. For further details, consult the oracle [164]§9.3.

[8]This is for the purposes of proof only. One does not need to do the SVD in practice.

and change basis via $\mathbf{P}$, writing $\mathbf{z} = \mathbf{P}\mathbf{z}'$ and $\hat{\mathbf{x}}_\mu = \mathbf{P}\mathbf{v}'$. Without loss of generality we may assume $\delta_1 \leq \delta_2 \leq \cdots \leq \delta_N$. In this notation $v_i' = (1+\mu\delta_i)^{-1} z_i'$. It is a point of interest that (in the limit $N \to \infty$)

$$\begin{aligned} \text{Arithmetic mean of } (\delta_i) &\to 2N(\mathbf{a}) \\ \text{Harmonic mean of } (\delta_i) &\to 2C(\mathbf{a}), \end{aligned}$$

a result that arises from considering the definitions of $N(\mathbf{a})$ and $C(\mathbf{a})$; the factor of 2 occurs because $\mathbf{F}$ produces the forward and backward prediction errors.

E and C methods

In each case let the LHS of the constraint equation be called $g(\mu)$ and the RHS g_0. We are going to show that $g(\mu)$ is monotone-decreasing.

For case E we have

$$g(\mu) = |\mathbf{F}\hat{\mathbf{x}}_\mu|^2 = \sum_{i=1}^{N} \frac{\delta_i |z_i'|^2}{(1+\mu\delta_i)^2}.$$

which is clearly monotonically decreasing; specifically it decreases from $|\mathbf{F}\mathbf{z}|^2$ to 0 as μ goes from 0 to ∞.

For case C we have

$$g(\mu) = \frac{|\mathbf{F}\hat{\mathbf{x}}_\mu|^2}{|\hat{\mathbf{x}}_\mu|^2} = \sum_{i=1}^{N} \frac{\delta_i |z_i'|^2}{(1+\mu\delta_i)^2} \Big/ \sum_{j=1}^{N} \frac{|z_j'|^2}{(1+\mu\delta_j)^2}.$$

Let the numerator be U and the denominator V, and let their μ-derivatives be U', V'. We are to show that $VU' - UV' < 0$. Using suffix i in the U- and U'- summations and suffix j in the V- and V'- summations, we find after a little algebra that

$$VU' - UV' = \sum_{i,j=1}^{N} \frac{-2|z_i'|^2 |z_j'|^2 \delta_i(\delta_i - \delta_j)}{(1+\mu\delta_i)^3(1+\mu\delta_j)^3}$$

Then interchanging the suffices, adding the expressions and dividing by 2 :

$$VU' - UV' = \sum_{i,j=1}^{N} \frac{-|z_i'|^2 |z_j'|^2 (\delta_i - \delta_j)^2}{(1+\mu\delta_i)^3(1+\mu\delta_j)^3} \leq 0$$

as required. So along $[0,\infty)$ g decreases and these are the bounds:

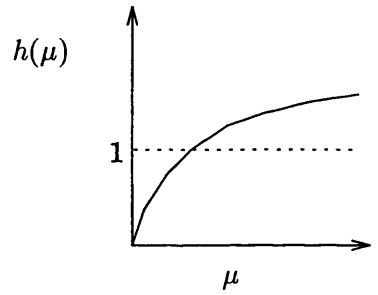
$$g(0) = |\mathbf{F}\mathbf{z}|^2 = \textstyle\sum_{i=1}^{N} \delta_i |z_i'|^2, \qquad g(\infty) = \dfrac{\sum_{i=1}^{N} |z_i'|^2/\delta_i}{\sum_{j=1}^{N} |z_j'|^2/\delta_j^2}. \tag{4.17}$$

R method

We are to solve the equation $\mu = h(\mu)$, where $h(\mu) = |\mathbf{z} - \hat{\mathbf{x}}_\mu|^2/|\mathbf{F}\hat{\mathbf{x}}_\mu|^2$. One way of doing this would be to iterate the equation, that is, to hope that the sequence $\mu, h(\mu), h(h(\mu)), \ldots$ converges and that there is only one stationary point. In the notation of the previous section,

$$\frac{h(\mu)}{\mu} = \sum_{i=1}^{N} \frac{\mu\delta_i^2|z_i'|^2}{(1+\mu\delta_i)^2} \Big/ \sum_{j=1}^{N} \frac{\delta_j|z_j'|^2}{(1+\mu\delta_j)^2}. \tag{4.18}$$

Clearly $\lim_{\mu\to 0} h(\mu)/\mu = 0$ and $\lim_{\mu\to\infty} h(\mu)/\mu = \infty$, inspiring the following sketch of $h(\mu)/\mu$:



A solution must therefore exist, and if the derivative of $h(\mu)/\mu$ is everywhere positive then the solution will be unique, in which case at that solution

$$0 < \frac{d}{d\mu}\frac{h(\mu)}{\mu} = \frac{\mu h'(\mu) - h(\mu)}{\mu^2} = \frac{h'(\mu)}{\mu} - 1$$

implying that $h'(\mu) > 1$. Therefore successive iterates $\mu, h(\mu), h^2(\mu), \ldots$ will *diverge* to 0 or ∞; this is seen in practice. Turning now to the question of monotonicity, in (4.18) write U for the numerator and V for the denominator; then proceeding as for the C and E cases we find

$$VU' - UV' = \sum_{i,j=1}^{N} \frac{\delta_i\delta_j|z_i'|^2|z_j'|^2[\delta_j + 3\mu\delta_i\delta_j - \mu\delta_j^2 + \mu^2\delta_i\delta_j^2]}{(1+\mu\delta_i)^3(1+\mu\delta_j)^3}.$$

After interchanging the suffices, adding the expressions, and dividing by 2, this becomes

$$\sum_{i,j=1}^{N} \frac{\delta_i\delta_j|z_i'|^2|z_j'|^2[\delta_i + \delta_j + \mu(6\delta_i\delta_j - \delta_i^2 - \delta_j^2) + \mu^2\delta_i\delta_j(\delta_i + \delta_j)]}{2(1+\mu\delta_i)^3(1+\mu\delta_j)^3}.$$

Unfortunately one cannot say that each summand is nonnegative; in fact there are a few cases where $h(\mu)/\mu$ undergoes a slight dip as μ increases, but in those cases the solution to $h(\mu)/\mu = 1$ was still unique. We have not found an example in which the R method failed to give a unique estimate $\hat{\mu}$.

Consideration of the coherent case

As with the estimation procedure, in which we proved that the generalised prediction error approach gave optimal results (on account of zero prediction error), it is worth considering what happens in the case of signal separation. It is important to realise that in this case the kernel of $\mathbf{F}$ (or equivalently of $\mathbf{F}^\dagger\mathbf{F}$) is nonempty and contains linear combinations of tones corresponding to the z-plane poles. Accordingly there will be p zero eigenvalues, $\delta_1, \ldots, \delta_p$.

Let us therefore consider how the three methods perform in this limit. The E case is straightforward, for $\sigma_\epsilon^2 = 0$ (so we require $g(\mu) = 0$) and $g(\mu) = 0 \Leftrightarrow \mu = \infty$ (so a solution exists). The C case is a little more complicated: using the fact that $\delta_1 = \cdots = \delta_p = 0$, we compute $g(\infty)$ using (4.17) and find it to be 0 (unless $z'_1 = \cdots = z'_p = 0$, but that would only happen if $\mathbf{z}$ contained no power at the frequencies we were trying to remove). For the C and E methods we therefore find the solution to be

$$v'_i = \begin{cases} z'_i, & \delta_i = 0 \\ 0, & \delta_i \neq 0 \end{cases}$$

which means that $\hat{\mathbf{x}} = \hat{\mathbf{x}}_\infty$ is the projection of $\mathbf{z}$ onto $\ker \mathbf{F}$. This means that the filter is performing a least-squares fit to the model $\sum_i A_i e^{i\Omega_i t}$, as one would hope.

The R method on the other hand does not have this property. Specifically, the equation $\mu = h(\mu)$ has a finite solution which does not tend to ∞ as the AR model becomes coherent. Consequently, in signal separations in which the AR process is harmonic the estimate $\hat{\mathbf{x}}$ of the harmonic process contains unwanted frequencies or, to put it differently, components which belong in $\hat{\mathbf{s}}$.

It therefore appears that the best of the three methods is type C (preferable to type E as it does not require the driving noise variance to be known). However we have found that when $\mu = \sigma_s^2/2\sigma_\epsilon^2$ is small ($\lessapprox 1$) the R method tends to provide a better estimate, particularly when the sampling is irregular (we shall discuss this next): the C method tends to underestimate μ and consequently produces an estimate of $\hat{\mathbf{s}}$ that is too small. Given that the R method underestimates μ in the coherent case (by producing a finite estimate when it should set $\hat{\mu} = \infty$) it appears that a neat way of combining the good points of each method is to estimate μ using both the C and R methods and take the higher of the two estimates. We call this the B method (B for 'both') and this is our method of choice.

4.5.3 Irregular sampling

To answer the second boxed question in §4.5.1 is now very easy. We simply replace the AR coefficients in $\mathbf{F}$ with the generalised forward and backward prediction coefficients:

$$\mathbf{F} = \begin{bmatrix} r_p^{p+1} & r_{p-1}^{p+1} & \dots & r_0^{p+1} & & & \\ & r_p^{p+2} & r_{p-1}^{p+2} & \dots & r_0^{p+2} & & \\ & & \dots & \dots & \dots & & \\ & & & r_p^N & r_{p-1}^N & \dots & r_0^N \\ s_0^{1*} & s_1^{1*} & \dots & s_p^{1*} & & & \\ & s_0^{2*} & s_1^{2*} & \dots & s_p^{2*} & & \\ & & \dots & \dots & \dots & & \\ & & & s_0^{N-p*} & s_1^{N-p*} & \dots & s_p^{N-p*} \end{bmatrix}$$

The normalisation described in §4.1 means that the rows of $\mathbf{F}$ are defined only up to multiplication by scalars of modulus 1. This is not a problem because only $\mathbf{F}^\dagger\mathbf{F}$ is used, and that *is* well-defined. So by design, all the above working holds good.

4.6 Test examples

4.6.1 Suppressing an interferer

We have considered a signal consisting of the broad-band process as defined in (4.12), with a (small) sinusoid added:

$$z(t) = y_c(t)\cos 2\pi f_0 t + y_s(t)\sin 2\pi f_0 t + A\cos 2\pi f_s t.$$

We can think of the broad-band process as an undesirable clutter signal (such as Bragg clutter [142] in HF radars) and the sinusoid as a 'ship target' appearing at a frequency governed by its radial velocity. The parameters of the broad-band process were the same as those considered in the spectral estimation examples (§4.4.3), i.e. $b' = 0.01\text{s}^{-1}$, $\sigma_y^2 = 50$, $f_0 = 0.21\text{Hz}$. The parameters of the added sinusoid were $A = 1$, $f_s = 0.61\text{Hz}$. First the signal was sampled (256 samples were taken) and the Lomb spectrogram calculated. From the spectral estimation work we know how to represent the broad-band signal as an AR(2) process, and the poles, together with the samples $z_n = z(t_n)$, were supplied to the filtering algorithm. The spectral content of the 'residual' ($\hat{\mathbf{s}}$) of the filter was then assessed, again using the Lomb spectrogram. The same sampling schemes as in §4.4.1 were applied.

Results are shown in Figure 4.9 for those six sampling schemes. Owing to the spectral leakage the sinusoid does not show up, except for regular sampling (in which it occurs twice as a consequence of aliasing). In the filtered signals, where the broad-band process has been removed, the sinusoid can be detected by elementary spectral analysis, as seen in the diagrams. Incidentally for sampling scheme (e), a jittered sampling scheme, the aliases are not suppressed very well; that jittered sampling is inferior to additive-random sampling, from the point of view of alias suppression, was discussed in §2.5.

Obviously the removal of the strong 'interferer' at 0.21Hz has greatly enhanced the detectability of the sinusoidal feature. As a quantitative guide to the improvement we shall describe a fairly obvious method of assessing the SNR of each spectral component in a signal consisting of sinusoids of known frequencies f_k in noise. It is a simple generalisation of the Lomb method. The time-domain signal is decomposed into two parts,

$$x(t) = \sum_k s_k(t) + r(t)$$

with

$$s_k(t) = A_k \cos 2\pi f_k t + B_k \sin 2\pi f_k t.$$

The weights A_k, B_k are found by SVD, minimising the residual $\sum_n |r(t_n)|^2$. The power P_k in each component, and the noise power, are defined as

$$\begin{aligned} P_k &= \frac{1}{N} \sum_{n=1}^{N} |s_k(t_n)|^2 \\ P_0 &= \frac{1}{N} \sum_{n=1}^{N} |r(t_n)|^2 \end{aligned}$$

and for each component the SNR may be written P_k/P_0. For the example we have considered there is only one spectral component, so the method is identical to the Lomb spectrogram. In later examples there will be several components.

Using these quantitative measures we have appended to each pair of spectra an assessment of the SNR before and after filtering, and it is fairly clear that the filtering works consistently for all the sampling schemes tried.

For a more complicated example we have considered the possibility of several 'targets'. We have used the same interferer as above (centre frequency 0.21) and added four frequency components of amplitude 2, $\sqrt{2}$, 2, 2 at frequency 0.41, 0.79, 1.05, 1.65 respectively. Results for sampling schemes (a) and (b) in §4.4.1 (regular and rectangular additive-random) are shown. These four diagrams (Figure 4.10) summarise the benefits and difficulties of irregular sampling. When the sampling is regular, the targets are visible both in the filtered and unfiltered signal, but as multiple copies because of aliasing. More seriously, a nasty accident befalls the second target (i.e. that at frequency 0.79), because it is indistinguishable from the interferer (at sampling rate 1) and is removed by the filter. When the sampling is random the targets cannot be seen in the raw spectrum, as a consequence of smearing; but they are clear in the filtered signal and do not appear as aliases. Of course there is no longer any special relationship between frequencies 0.21 and 0.79 when the sampling is random, so the second target is not annihilated by the filter.

4.6.2 Estimating an FSK signal in noise

In the previous examples the AR process, whose poles were known, was viewed as an undesirable interferer, and after the separation the estimate of it ($\hat{\mathbf{x}}$) was discarded and the residue ($\hat{\mathbf{s}}$) analysed. Here we shall consider an example in which the AR process is the one in which we are interested, and the uncharacterised residual is

observation noise. This can therefore be viewed as an example of noise reduction. See Figure 4.11.

Frequency shift keying is a simple method of transmitting, by frequency modulation, a stream of digits[9]. For each type of digit a frequency is assigned and the transmitter is switched between the appropriate frequencies to transmit the information. The waveform is seen to be the solution of the differential equation

$$\frac{dx(t)}{dt} = 2\pi \mathrm{i} f_{c(t)} x(t), \qquad c(t) = c_{\lfloor t/T \rfloor}$$

in which $c_0, c_1, \ldots$ is the sequence of digits to be transmitted, T is the dwell (i.e. transmission time allotted to each digit) and $\lfloor \cdot \rfloor$ denotes the integer part. It is apparent that during the transmission of any one digit the signal is one of a selection of pure tones, and hence that the sampled signal obeys a coherent AR model with poles $\alpha_i = e^{2\pi \mathrm{i} f_i \delta t}$. Of course the signal in its entirety is not exactly predictable, because a prediction error occurs when the signal is switched from one frequency to another; however, as the waveform is a continuous[10] function these prediction errors will be quite small (and tend to 0 as $\delta t \to 0$). An AR model is likely to be appropriate, and in the binary case, when there are only two frequencies, we have a complex AR(2) model.

As a specific example we have used 0.10Hz and 0.19Hz as the transmission frequencies, 19s as the dwell, and an amplitude of 1. The transmitted code is 01010011100$\cdots$. The real and imaginary parts of this signal are shown (Figure 4.11(a)). An AR(2) model was then fitted to the signal using regular sampling at rate 1Hz; the AR(2) spectrum is shown (Figure 4.11(b)).

Next, samples were taken of the FSK signal using the sampling schemes (a) and (b) of §4.4.1 and complex white Gaussian noise of variance 1 was then added. This means that in each case the SNR, defined as the FSK signal power divided by noise power, is 0dB. Figure 4.11(c,e) show the noisy signals, and the transmission pattern is not at all obvious. Then the filtering algorithm was employed to separate the AR model (clean FSK signal) from the noise. The signal estimates ($\hat{\mathbf{x}}$) are shown in Figure 4.11(d,f). In each case the transmission pattern is much clearer; it could, for example, be picked out by examining the zero-crossings.

While on the subject of noise reduction, we may usefully point out similarities with noise reduction in nonlinear dynamics. There the idea is, given a 'chaotic' signal that we know how to predict, to clean it up if noise is added to it. There are similarities and differences between the techniques used here and the nonlinear ones, which we now discuss from the point of view of regular sampling. In the nonlinear case one has

$$x_n = H(x_{n-1}, \ldots, x_{n-p})$$

and let us assume that H is known (or that H has been estimated from the data in

[9]This is for the purposes of demonstrating a principle and the modulation scheme discussed here should not be regarded as state-of-the-art. For a comprehensive discussion of digital communication signals, see [167].

[10]i.e. 'not discontinuous', as opposed to 'not discrete'

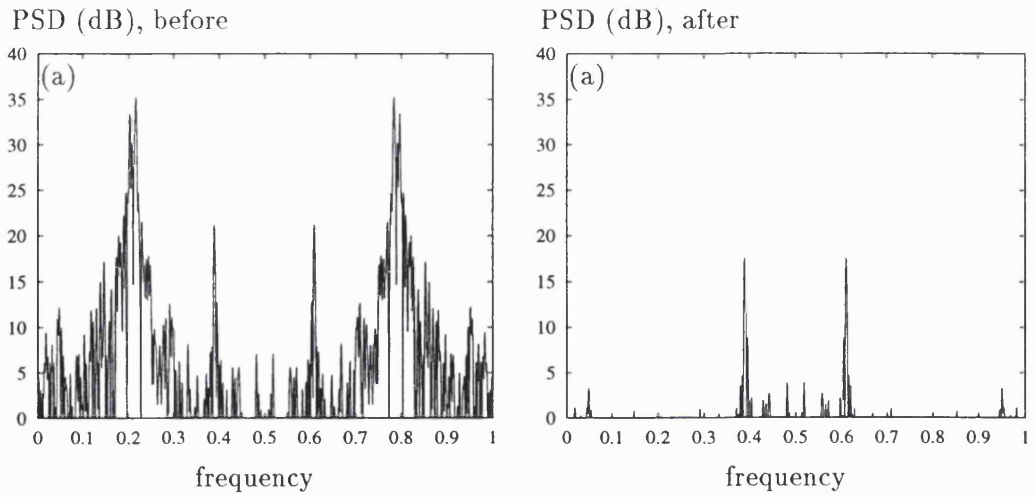
embedding space). The idea is then to minimise the total dynamical error

$$E_{\text{dyn}} = \sum_{n=p+1}^{N} |\xi_n - H(\xi_{n-1}, \ldots, \xi_{n-p})|^2 ;$$

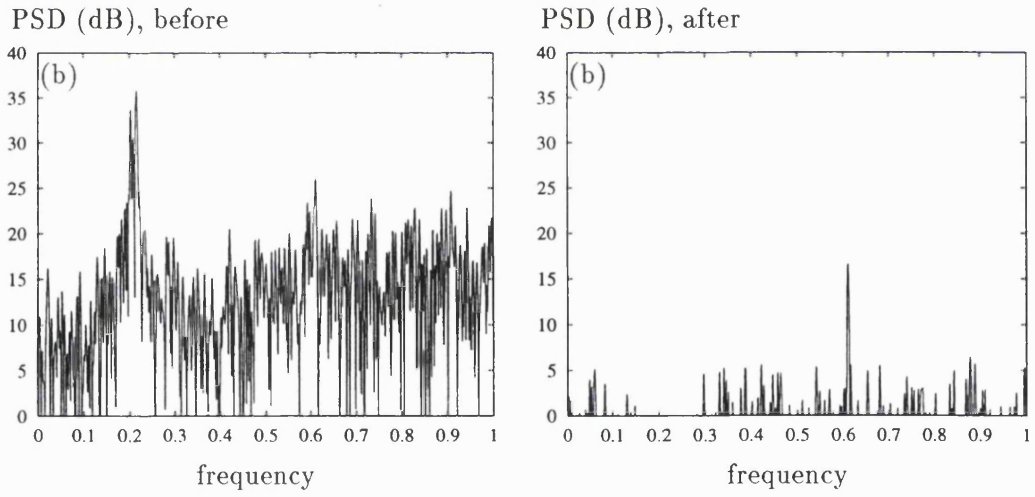
this requires a nonlinear minimisation, and hence is usually tackled by gradient descent or a similar technique ([52]; see also [51]). As there will be many trajectories (specifically, a p-dimensional set) that minimise E_{dyn}, the objective is to find one that is close to the original data, and hence one can write down an objective function of the form

$$\lambda \sum_{n=p+1}^{N} |\xi_n - H(\xi_{n-1}, \ldots, \xi_{n-p})|^2 + \sum_{n=1}^{N} |z_n - \xi_n|^2$$

where $\mathbf{z}$ is the uncleaned time series, and λ is a Lagrange multiplier. It is not difficult to see that this is the same type of idea used in the construction of the Wiener filter for the linear case. Note however that in the linear case we do not need to know the amplitude of the signal $\mathbf{x}$, whereas in the nonlinear case we need either to know it or to be able to estimate it from the data (this amounts to finding H). Note also that a linear function H gives rise to a linear optimisation problem, because then even when the sampling is irregular the objective function is quadratic in $\boldsymbol{\xi}$.

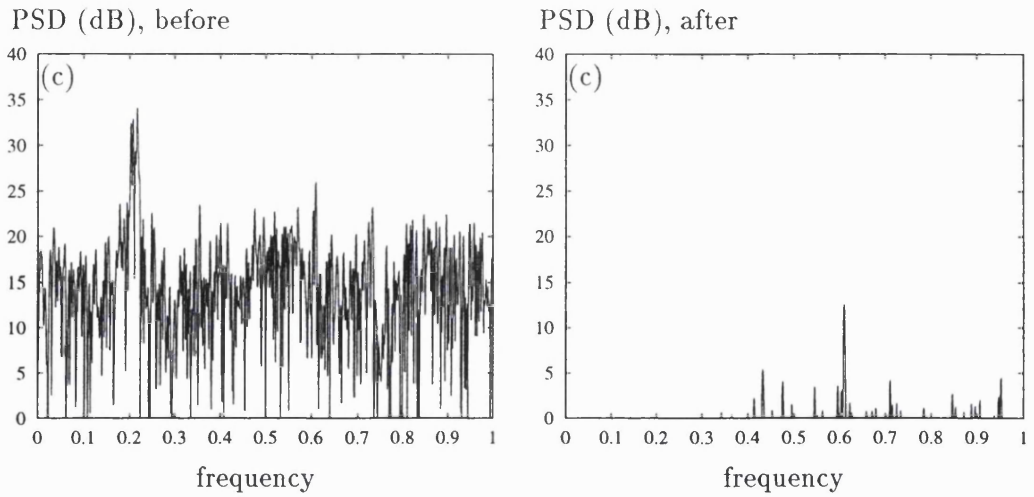


Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	21.3	18.9	−17.7	+0.8	+18.5
Others	39.0	18.1			

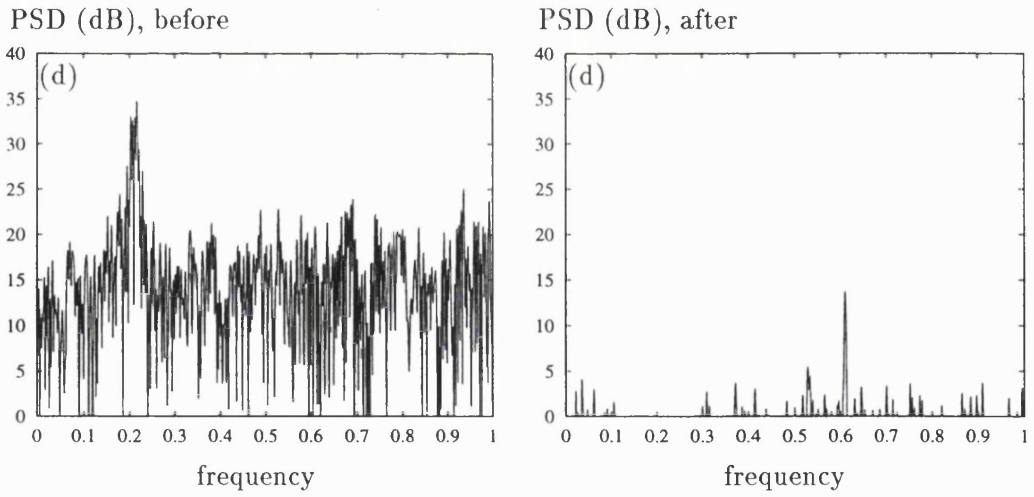


Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	24.3	18.5	−15.2	−1.3	+13.9
Others	39.5	19.8			

Figure 4.9. Effect of filtering (sampling schemes a,b from §4.4.1). The target is at frequency 0.61; the interferer is 'centred' at 0.21.

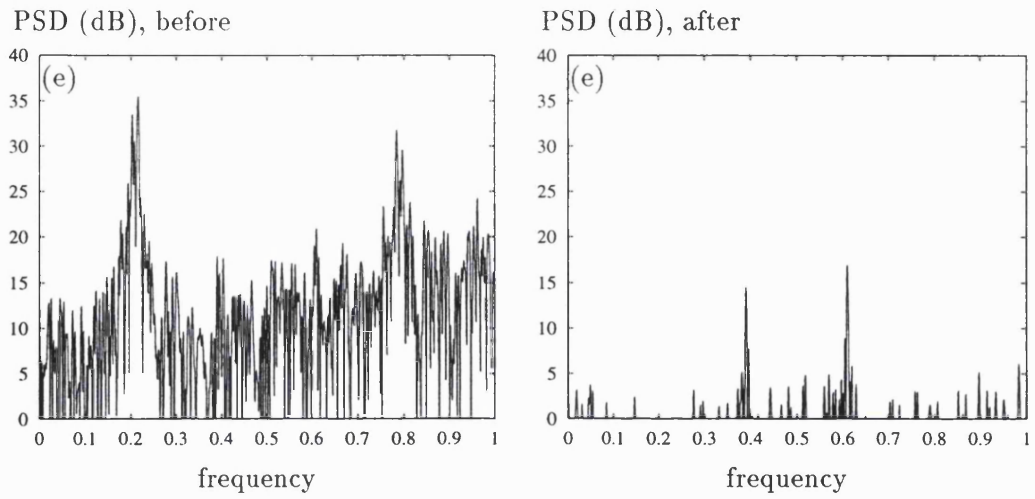


Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	24.2	13.8	−14.2	−3.0	+11.2
Others	38.4	16.8			

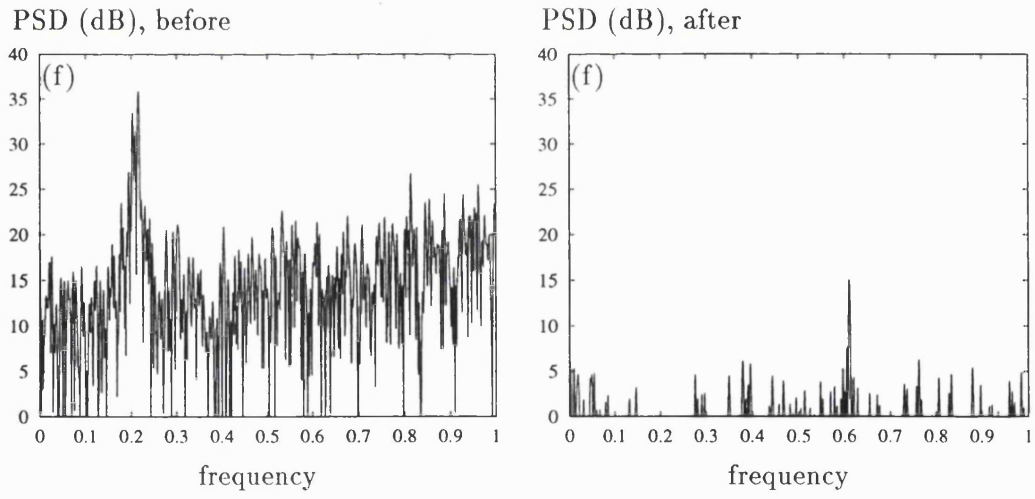


Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	20.5	15.1	−18.8	−2.9	+15.9
Others	39.3	18.0			

Figure 4.9. Effect of filtering (sampling schemes c,d from §4.4.1). The target is at frequency 0.61; the interferer is 'centred' at 0.21.

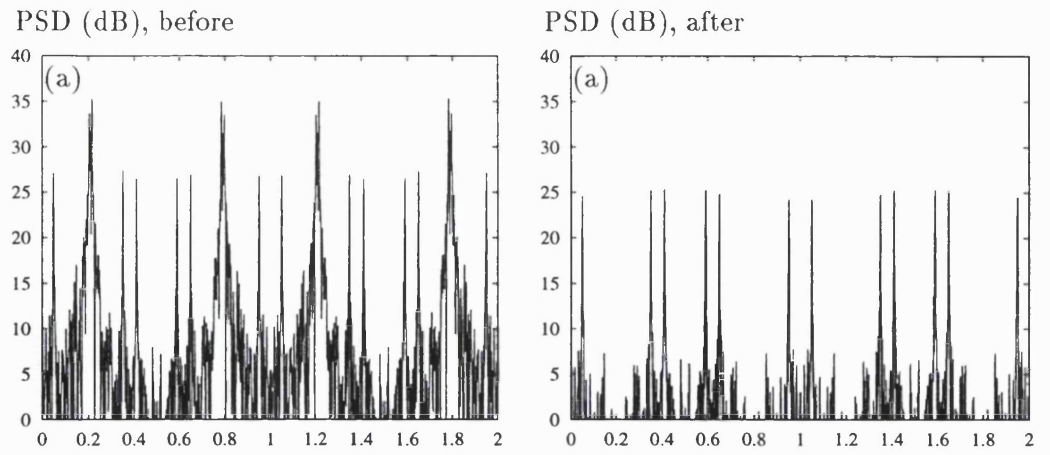


Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	21.0	18.5	−18.2	−0.6	+17.6
Others	39.2	19.1			

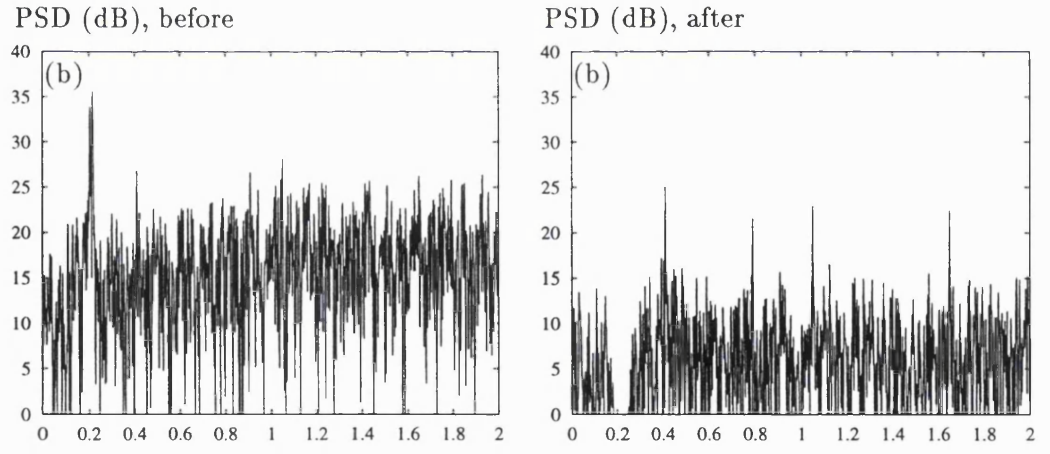


Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	21.2	17.1	−18.3	−2.9	+15.4
Others	39.5	20.0			

Figure 4.9. Effect of filtering (sampling schemes e,f from §4.4.1).
The target is at frequency 0.61; the interferer is 'centred' at 0.21.



Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.41	26.7	25.4	−12.3	+5.0	+17.3
0.79	27.8	−11.1	−11.2	−31.5	−20.3
1.05	27.3	24.6	−11.7	+4.2	+15.9
1.65	27.5	25.2	−11.5	+4.8	+16.3
Others	39.0	20.4			



Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.41	27.0	25.3	−12.4	+2.1	+14.5
0.79	22.2	20.0	−17.2	−3.2	+14.0
1.05	27.4	22.4	−12.0	−0.8	+11.2
1.65	25.3	21.8	−14.1	−1.4	+12.7
Others	39.4	23.2			

Figure 4.10. Effect of filtering, four targets: (a) uniform, (b) random sampling. The targets are at 0.41, 0.79, 1.05, 1.65; the interferer is 'centred' at 0.21.

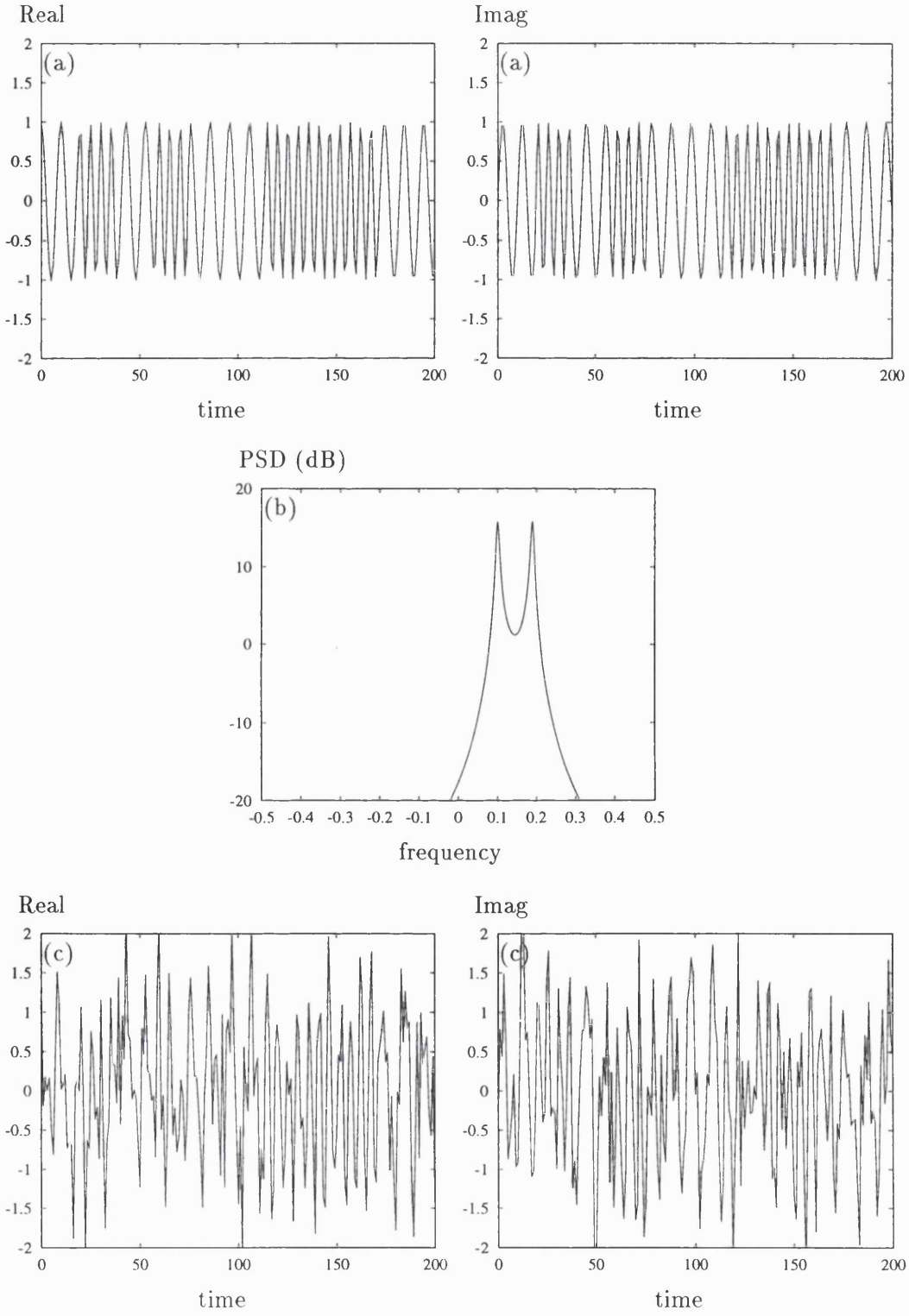


Figure 4.11. Modelling and estimation of FSK signal in noise (see overleaf).

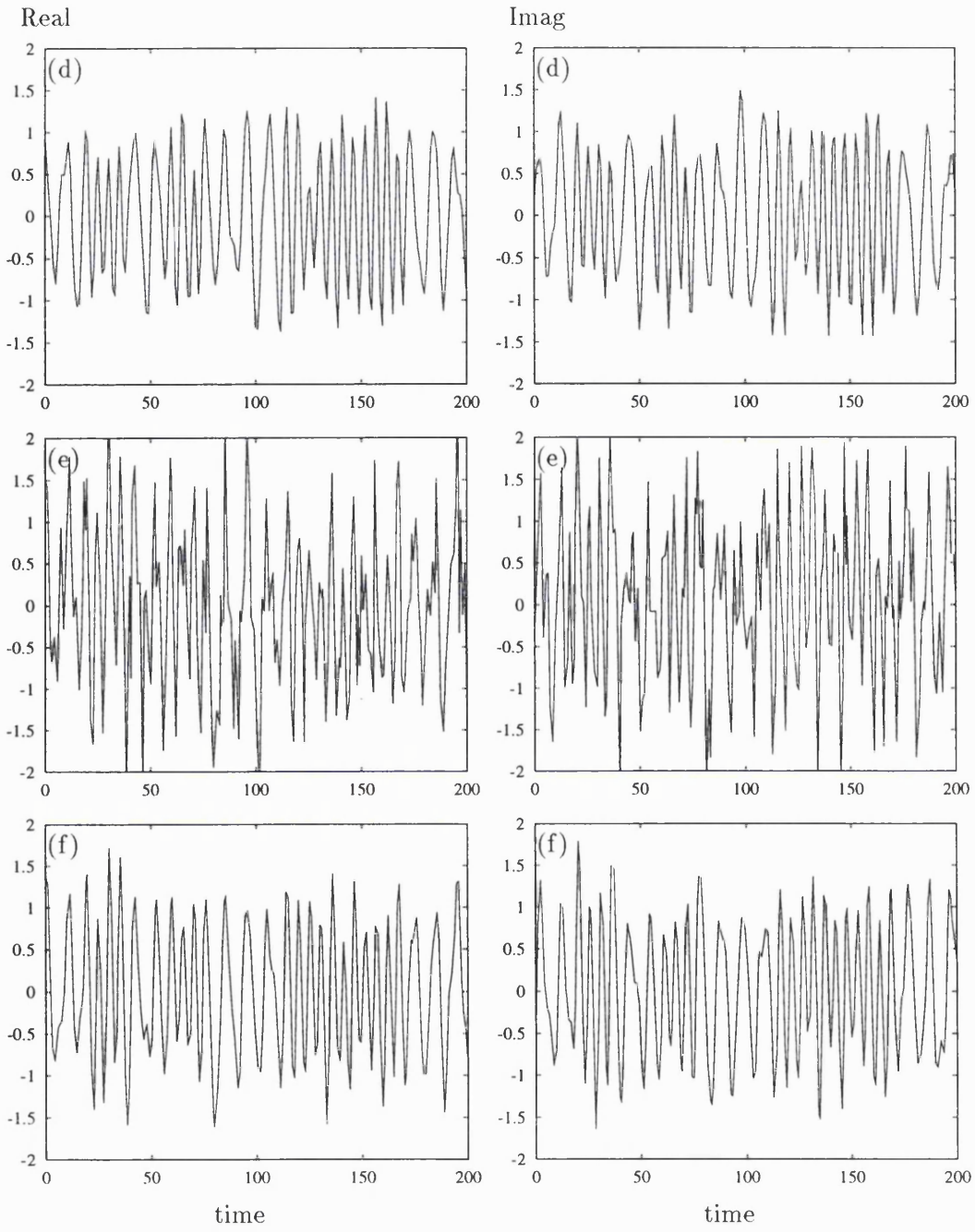
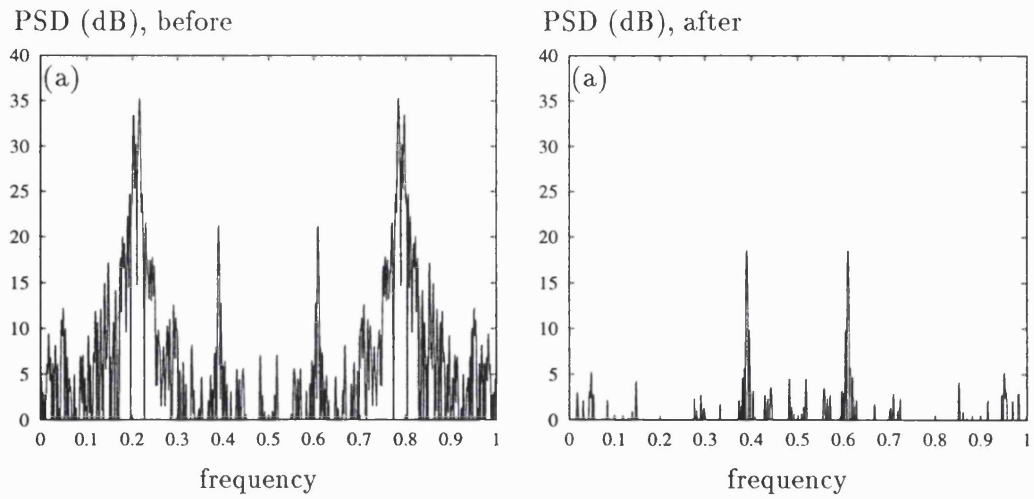


Figure 4.11 (second page)

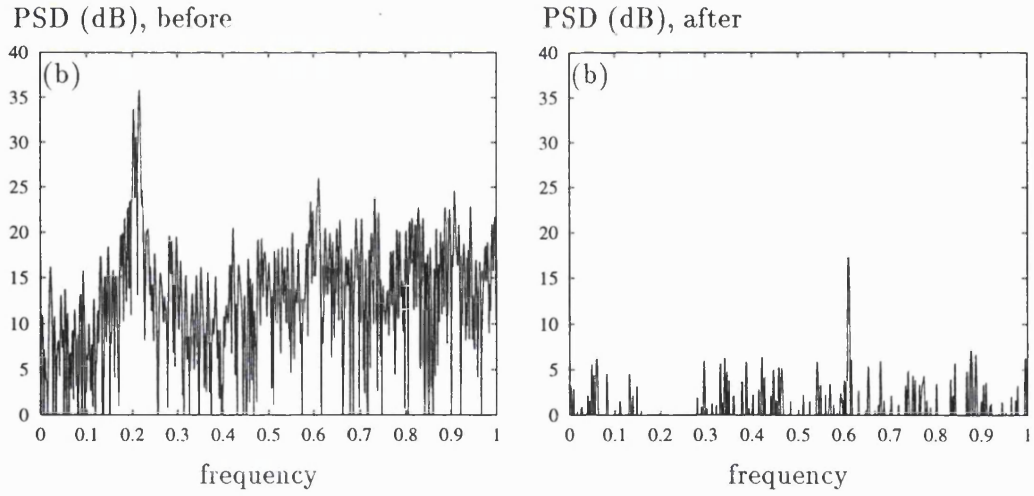
(a) Clean FSK signal (real & imag.) (b) AR(2) model of clean FSK signal.
(c) Noisy signal, regular sampling (d) Cleaned signal, regular sampling
(e) Noisy signal, irregular sampling (f) Cleaned signal, irregular sampling

4.7 The identification problem (revisited)

The problem of identifying the spectral components of an irregularly sampled signal has led in the astronomy literature a sequence of algorithms, offering successively better performance. Indeed it seems *de rigueur* to name them after their supposed cleanliness; naturally we refer the reader only to the cleanest [69]. These and related cleaning algorithms [20] revolve around the simple principle of estimating the strongest frequency components using the periodogram and then removing them in the time domain before reassessing the spectrum. Of course these methods simply correspond to the coherent case discussed in §4.5.2, wherein it was shown that performing a least-squares fit to a sinusoid (or sinusoids) is the same as filtering using our algorithm with AR poles on the unit circle and letting $\mu \to \infty$ (as prescribed by the C estimation scheme). Our technique will deal with more general classes of signals, thereby including the case in which the amplitude and/or frequency of the sinusoid wander over a period of time. It is clear that the AR approach, which essentially relies on local linear predictability, is more flexible than the current approach of using a global representation $A \cos \Omega t + B \sin \Omega t$. In other words, the AR approach admits spectra other than delta-functions, i.e. spectra having nonzero bandwidth. All we have to do is to combine the model-fitting algorithm with the filtering one. As a test we repeat the tests of §4.6.1 but instead of knowing in advance the poles of the AR process to be removed, we estimate them from the data (which contains the extraneous sinusoid as well). It is to be hoped, of course, that the presence of the sinusoid will not drastically affect the estimation process. We show the results for sampling schemes (a) and (b) only: as can be seen, the results are virtually identical with those of §4.6.1 where the model poles were estimated from the 'clean' interferer.



Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	21.3	18.5	−17.7	+0.8	+18.5
Others	39.0	17.7			



Frequency	Power (dB)		SNR (dB)		
	Before	After	Before	After	Gain
0.61	24.3	17.4	−15.2	−1.1	+14.1
Others	39.5	18.5			

Figure 4.12. Effect of filtering (sampling schemes a,b). In this simulation the process paramaters are estimated from the test data, rather than being known beforehand.

4.8 Final remarks

We have developed the theory of linear prediction for irregularly sampled data that relies on linking discrete-time process with the continuous-time autoregressive and harmonic models. This has allowed estimation and filtering algorithms to be constructed, and there is plenty of scope for research in either of these areas.

One area that could be investigated—though it is by no means easy—is the construction of the function E, as the normalisation of the predictors is an awkward sticking-point. Our normalisation (4.4) has some useful properties (notably it simplifies the analysis of prediction errors for a white input signal) and on the basis of the test results we can say that it works well—the peak heights of the ARPSDs in Figures 4.7–8, which are extremely sensitive to small errors in estimating the poles, are remarkably consistent—but a sounder theoretical justification would be helpful. The other awkward corner is that a continuous-time AR process when sampled (regularly or irregularly) produces a discrete ARMA$(p, p-1)$ process, not an AR(p) process, and it is not clear whether, or how, to take into account the MA$(p-1)$ part.

The optimisation to obtain $\hat{\boldsymbol{\alpha}}$ from $E(\mathbf{x}, \boldsymbol{\alpha})$ is nonlinear because each residual does not depend linearly on the (α_i). This is rather inconvenient, and means that one always has the worry about finding the best minimum. There does not seem to be a way out of that problem. However we have found that, if one is prepared to make a quick search of the performance surface and choose the best point as the starting-value, there is little difficulty in obtaining reproducible results. Remember that in the regularly-sampled case the performance surface is quadratic in the AR coefficients, and hence unimodal when viewed as a function of the poles (we mean unique up to a permutation of the poles). Incidentally Belcher *et al.* [12] reconsidered Jones' work [101, 102] and came up with a reparametrisation of the Laplace-domain (continuous-time AR) poles to circumvent the $p!$-to-1 mapping from poles to coefficients.

While on the subject of nonlinear optimisations, we note that there is one minor improvement that can be worked in at no extra computational cost, namely the use of methods from *robust statistics* [89, 164] as an alternative to least-squares. Least-squares methods fall down when outliers in the data give rise to excessive contributions on squaring and pull the estimated parameter away from its correct value; they can be attributed to the tails of the Normal distribution, which are unrealistically thin. Robust statistics assumes that the underlying distribution ϕ of the residuals (e_n) is non-Gaussian (with fatter tails); in particular

$$L = \ln \mathrm{lik}(\mathbf{e}|\boldsymbol{\alpha}) = \ln \prod_n \phi(e_n) = \sum_n \ln \phi(e_n)$$

is maximised. In linear least-squares problems e_n is linear in $\boldsymbol{\alpha}$ and if $\ln \phi$ is quadratic (as it is for a Gaussian ϕ), then $\partial L/\partial \boldsymbol{\alpha}$ is linear in $\boldsymbol{\alpha}$. As soon as one moves away from Gaussian distributions, $\partial L/\partial \boldsymbol{\alpha}$ becomes nonlinear, which is inconvenient. For our application e_n is not linear in $\boldsymbol{\alpha}$ to start with, so the introduction of further nonlinearities is not of much concern. Another possibility is to weight the prediction errors in inverse proportion to their associated intersample spacings, for if the data

had a few large gaps one would not expect to be able to estimate across the gap, and reducing the sensitivity of $E(\mathbf{x}, \boldsymbol{\alpha})$ to these terms would give better robustness.

Concerning the filtering, there are several questions of interest. One is the relationship between the C and R methods of estimating the variance ratio μ, and why they give different results. As a third method there is also the expectation-maximisation (EM) algorithm [151] for estimating this parameter. There is also the question of how to implement the filtering algorithm for data that arrive sequentially rather than in blocks; one possibility is to solve the filtering equation for blocks of data $(t_n, x_n)_{n=1}^{N}$ and $(t_n, x_n)_{n=2}^{N+1}$ and find a relation between the two to give the required recursion. In the case of sequential component extraction (using alternate estimation and filtering) it would be worth investigating whether, if several components are identified in a signal, they should be extracted sequentially or simultaneously.

Given the vast literature on autoregressive methods, the ideas introduced here could form the basis of many interesting developments. Further there is no reason why one should stick to simple autoregressive models. In recent years there has been much interest in allowing the parameters of an AR model to be amplitude-dependent, or to consider other forms of nonlinearity such as the *bilinear model* of which the following is the discrete-time representation:

$$y_n + \sum_{j=1}^{p} a_j y_{n-j} = e_n + \sum_{i,j=1}^{r} y_{n-i} e_{n-j}.$$

Priestley [166] discusses these nonlinear models in discrete time. The continuous-time *threshold* autoregressive model, or CTAR, is discussed in [27, 213, 91].

Chapter 5

Nonlinear models

In this chapter we shall consider nonlinear prediction of a time series sampled at irregular intervals. The motivation for nonlinear prediction is the study of signals obtained from observations of a dynamical system. In the regularly sampled case we can predict an observation from d previous observations provided that the embedding dimension d is sufficiently large. In the irregular case we expect to have to take the intersample spacings into account, so that the nth sample is predicted from the previous d samples and their associated intersample spacings. Recent research [198, 202] shows that this approach will in principle work. The simplest nontrivial case is that of a sinusoidal signal, which can be viewed as one-dimensional dynamics $\dot{\theta} = \Omega$, on a circle parametrised by the angle θ, using an observation function $g : \theta \mapsto \cos\theta$. It turns out that even this simple case is quite interesting when the sampling is irregular, and the existence of a smooth predictor function depends on a certain condition relating the frequency of the sinusoid and the irregularity of the sampling. Then we consider a rather different type of dynamical example, the Lorenz attractor, in which chaotic dynamics are predicted. Chaotic sources generally give rise to broad-band time series which are not analysed well using linear techniques. The nonlinear techniques shown here give a significant improvement. We conclude by suggesting how methods for nonlinear prediction demonstrated here, when combined with the linear filtering discussed in Chapter 4, could be used to design nonlinear filters for irregularly sampled data. Thus nonlinear prediction and filtering techniques fit in with the general aim in this thesis of devising methods for analysing irregularly sampled time series.

5.1 Predicting irregularly sampled data

In §2.7 we discussed Takens' theorem and nonlinear prediction for the case of regular sampling and said that:

- Suppose that a signal is obtained by observing, using a smooth function g, dynamics $y(t)$ on a compact manifold $\mathcal{M}$ of dimension D. Construct the set E of d-dimensional delay-vectors

$$\mathsf{E} = \left\{ [\, g\phi^{-1}y \quad g\phi^{-2}y \quad \cdots \quad g\phi^{-d}y \,]^{\mathsf{T}} : y \in \mathcal{M} \right\}$$

where ϕ takes $y \in \mathcal{M}$ to where y is moved by the dynamics after time τ. We have a copy of $\mathcal{M}$ in E provided $d \geq 2D+1$. In fact $\mathcal{M}$ and E are identical up to a smooth coordinate change. We are taking the genericity constraints for granted.

- Time-discretised dynamics on $\mathcal{M}$ are manifested in E. Thus there is a smooth function $K : \mathsf{E} \to \mathsf{E}$ that predicts the $(n+1)th$ delay-vector from the nth. Writing $x_n = gy(t+n\tau)$ and $\underline{x}^n = [\, x_{n-1} \quad x_{n-2} \quad \cdots \quad x_{n-d} \,]^\mathsf{T}$, so that $\underline{x}^n \in \mathsf{E}$,

$$K : \underline{x}^n \mapsto \underline{x}^{n+1}, \text{ and } K : \mathsf{E} \to \mathsf{E} \text{ is smooth.}$$

- We don't need all the components of K, as $(d-1)$ of them are trivial. The important thing is to predict x_n (the 'top' component of $\underline{x}^{n+1}$) from $\underline{x}^n$:

$$H : \underline{x}^n \mapsto x_n, \text{ and } H : \mathsf{E} \to \mathbf{R} \text{ is smooth.}$$

A good way to estimate H, from a set of data, is the radial basis function method, using known delay-vectors as centres. Thus,

$$x_n \approx \hat{H}(\underline{x}^n) = \sum_j \lambda_j \psi(\|\underline{x}^n - \underline{x}^{c_j}\|).$$

The (c_j), which are just integers, index the centres.

When the sampling is irregular, the generalised version of Takens' theorem, due to Stark, Broomhead *et al.* [198, 202], states that:

- H still exists, if we incorporate the intersample spacings into the delay-vectors, and is well-behaved except at countably many points.

So one question to be answered is whether the 'exception at countably many points' occurs in practice and, if it does occur, what the implications are for time series embedding and prediction. It is then a question of how to approximate H. Again we wish to use radial basis functions. To do this, write $\tau_n = t_{n+1} - t_n$ and define delay-vectors by

$$\underline{\underline{x}}^n := \left[\, \underline{x}^n \quad \underline{\tau}^n \,\right]^\mathsf{T} = \begin{bmatrix} x_{n-1} & x_{n-2} & \cdots & x_{n-d} \\ \tau_{n-1} & \tau_{n-2} & \cdots & \tau_{n-d} \end{bmatrix}^\mathsf{T}. \tag{5.1}$$

We can put the usual algebraic structure on these; the norm is given by

$$\|\underline{\underline{x}}\|^2 \equiv \left\| \underline{x} \quad \underline{\tau} \right\|^2 := w_x \|\underline{x}\|^2 + w_\tau \|\underline{\tau}\|^2 \tag{5.2}$$

To make this norm dimensionally consistent w_x and w_τ will need to have different dimensions; in fact, as there are two parameters in the above equation, ψ need not have an implied 'width parameter'. We write $\hat{H}$ for the RBF approximant to the smooth function taking $\underline{\underline{x}}^n$ to x_n :

$$\hat{H} : \underline{\underline{x}}^n \mapsto \sum_j \lambda_j \psi(\|\underline{\underline{x}}^n - \underline{\underline{x}}^{c_j}\|).$$

We intend to use the Gaussian $\psi(r) = e^{-\frac{1}{2}r^2}$ throughout, though other choices such as $\sqrt{1+r^2}$ and $r^2 \ln r$ are possible too.

Before applying this we need to estimate w_x and w_τ, and after nondimensionalising they may be cast in the form

$$w_x = \frac{w'_x}{\mathrm{Var}\{x\}}, \quad w_\tau = \frac{w'_\tau}{\mathrm{Var}\{\tau\}}, \tag{5.3}$$

where $\mathrm{Var}(\cdot)$ denotes the variance (though other suitable measures of spread could be used) and $w'_{x,\tau}$ are dimensionless. This means that the parameters $w'_{x,\tau}$ depend on the nature of the problem rather than on the dimensions used to describe it. For the problems that we have been examining, $w'_x = 1$ and $w'_\tau = \frac{1}{25}$ were always found to be suitable.

Having selected $w'_{x,\tau}$ for a given data set, we must now decide which points in the training set should be used as centres. From these data points we wish to choose a representative set of centres that are not too close (numerical ill-conditioning will result if they are). As discussed in the Introduction, a simple way of accomplishing this is to start selecting centres at random from the time series and, each time a centre is chosen, reject it if it is too close to any of the previously chosen centres. We suggested in §2.7 that for the Gaussian RBF $f : \underline{x} \mapsto e^{-\frac{1}{2}w\|\underline{x}-\underline{c}\|^2}$, for which $f(\underline{c}) = 1$, a centre should be rejected if the value of its associated basis function at any of the previously chosen centres exceeds 0.9. We shall use the same criterion here.

Then, taking the training data $(x_n, t_n)_1^N$, we have only to minimise the following function with respect to the (λ_j) :

$$\text{Error} = \sum_{n=d+1}^{N} \left(x_n - \sum_j \lambda_j \psi(\|\underline{\underline{x}}^n - \underline{\underline{x}}^{c_j}\|) \right)^2 = \|\mathbf{A}\boldsymbol{\lambda} - \mathbf{b}\|^2$$

where $A_{ij} = \psi(\|\underline{\underline{x}}^i - \underline{\underline{x}}^{c_j}\|)$ and $b_i = x_i$. This is a completely standard linear least-squares problem: methods for solving it were discussed in §2.7. We shall use SVD on the matrix $\mathbf{A}$ to accomplish this; the tolerance level in the SVD will be 10^{-4} throughout.

5.2 Prediction of a sinusoid

The sinusoidal signal $\cos 2\pi ft$ contains the ingredients of an interesting dynamical prediction problem. It can be regarded as the observation of the system $\dot{\theta} = \Omega$ on a circular manifold $\mathcal{M} = \mathbf{T}^1$ with the (2-to-1) observation function $g : \theta \mapsto \cos\theta$. Note that $\mathcal{M}$ is compact. Alternatively we may view the dynamics in $\mathbb{R}^2(\xi, \eta)$ and say that we are observing the ξ-coordinate of the system ($\dot{\xi} = -\eta$, $\dot{\eta} = \xi$); there is nothing wrong with that, but it is a rather loose representation and hides the topology.

Consider, for an average sampling rate of 1 (frequency units), the prediction of a sinusoid of frequency f sampled using an additive-random sampling scheme with intersample spacings taken from the rectangular distribution with pdf

$$p(\tau) = \begin{cases} q^{-1}, & |\tau - 1| \leq \frac{1}{2}q \\ 0, & \text{otherwise.} \end{cases}$$

Additive-random sampling means that the intersample spacings are independent. It is apparent that:

- When $q = 0$, prediction will be accurate using $d = 2$ regardless of f because we know that a sinusoid can be predicted from two equispaced previous values by *linear* prediction, so the (more general) nonlinear prediction must also work. [The prediction from regular samples is $x_n = (2\cos 2\pi f\delta t)x_{n-1} - x_{n-2}$.]
- When f is small, prediction will be accurate for $q > 0$. This is because the signal will vary only slowly between the samples, so there should not be any difficulty in predicting it.

So the question is whether accurate reconstruction requires some constraint on f and q.

In an extensive numerical simulation we investigated the predictability of a sinusoid, using d previous values and associated intersample spacings, as a function of the sinusoid's frequency (f) and of the irregularity of the sampling (q). Both q and f were varied between 0 and 1. A sample size of 2000 was used. The centres (c_j) were chosen between $d + 1$ and 1000. The first 1000 data points were used for training, i.e. a linear least-squares fit was performed to these data points to obtain an estimate of the (λ_j). The second 1000 data points were used for verification. The mean-square verification error e_{ms} and the error in decibels e_{dB} were calculated in the natural way:

$$\begin{aligned} e_{\text{ms}} &= \sum_{n=d+1}^{N} |x_n - \hat{H}(\underline{x}^n)|^2 \Big/ \sum_{n=d+1}^{N} |x_n|^2 \\ e_{\text{dB}} &= 10\log_{10} e_{\text{ms}}. \end{aligned}$$

Here the data points indexed by $(x_n)_1^N$ are the verification data (in this demonstration, samples 1001 to 2000 of the time series).

In the first instance a model order of 2 (i.e. $d = 2$) was chosen, and the results are shown in Figure 5.2(a,b) for 30 radial basis centres and also for 60 centres. Although

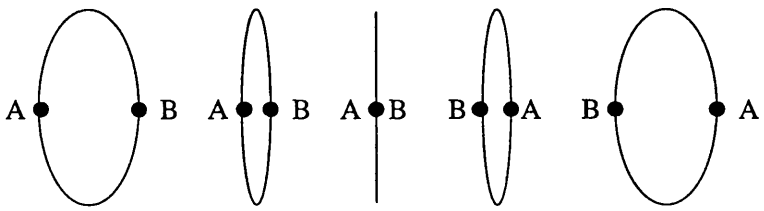
the latter allows a closer fit it is clear that there are some regions of the (q, f) plane that do not permit prediction. This is made clearer in Figure 5.2(c,d), which are contour plots of the three-dimensional graphs (a,b) showing lines of constant prediction error. A simple conclusion would be that when q or f is small, accurate prediction is possible, but not otherwise. This would, however, ignore the rather interesting 'kink' in the surface which is observed when q is kept at a low value and f is raised from 0.5 to 0.8 : the mean-square prediction error, having risen to a peak at around $f = 0.5$, drops and then increases again. The significance of $f = 0.5$ is that it is half the average sampling rate, but beyond that more analysis is required.

It turns out that the stumbling-block that prevents nonlinear prediction in the 'disallowed' regions of the (q, f) plane is precisely the same one that caused the normalisation problem in Chapter 4. The reader might recall that that difficulty was caused by trying to predict the observation at time t_n from observations t_{n-1}, t_{n-2} close to the zero-crossings of the sinusoid. It was resolved by replacing a classical prediction error with a generalised prediction sum $\sum_{j=0}^{d} r_j^n x_{n-j}$ and constraining the length of the coefficient vector $\mathbf{r}^n$. In the nonlinear prediction schemes that we are examining, we have returned to the prediction of x_n from two previous values—and so the difficulty presents itself again.

The dependence of x_n on previous samples is

$$x_n = \frac{\sin \Omega(t_n - t_{n-2})}{\sin \Omega(t_{n-1} - t_{n-2})} x_{n-1} - \frac{\sin \Omega(t_n - t_{n-1})}{\sin \Omega(t_{n-1} - t_{n-2})} x_{n-2}, \qquad \Omega = 2\pi f.$$

This function is singular whenever $2\pi f(t_{n-1} - t_{n-2})$ is a multiple of π. Nonlinear prediction schemes require the prediction function to be smooth, and so if the dependence of $H(\underline{\underline{x}})$ on $\underline{\underline{x}}$ is discontinuous, the method will fail. This situation can be imagined in embedding space. In the sketch below, the dynamics are embedded for different sets of intersample spacings.



In the first and last sketches the circular dynamics are embedded as an ellipse, but in the awkward case in the middle sketches, the minor axis of the ellipse is very narrow (and zero in the singular case). The implication for prediction is that the prediction function H has to take very different values at the points A and B, because at those points the dynamics are going in opposite directions. Consequently an attempt to approximate H using smooth functions (such as RBFs) will fail if A and B are very close.

The above discussion points to a conclusion that prediction is likely to fail if $\sin \Omega(t_{n-1} - t_{n-2})$ is ever zero (equivalently, if $2(t_{n-1} - t_{n-2}) \in \mathbb{Z}$). Accordingly, it appears that successful prediction of a sinusoid from two previous observations requires the set $\{2\tau_n f\}$ to lie completely within one connected component of the set

$\mathbf{R}\backslash\mathbf{Z}$. For the sampling scheme that we are considering, i.e. an additive-random one in which τ lies between[1] $(1 \pm \frac{1}{2}q)\bar{t}$, the allowable regions of the (q, f) plane are those for which $(2 \pm q)\bar{t}f$ both lie between k and $k+1$ for some integer k. This statement can be summarised by saying that the allowable region is $\mathcal{A}$ given as the following disjoint union:

$$\begin{aligned} \mathcal{A} &= \bigcup_{k \geq 0} \mathcal{A}_k \\ \mathcal{A}_k &= \{(q,f) : 0 \leq q \leq 2,\ f > 0,\ (2-q)\bar{t}f > k,\ (2+q)\bar{t}f < k+1\} \end{aligned} \tag{5.4}$$

The $\mathcal{A}_k$ are shown in Figure 5.1 and the allowed regions are also marked on the test results (Figure 5.2(e)). It appears that the contour plots of prediction error closely follow the allowed regions, so the 'kink' in the performance surface is explained.

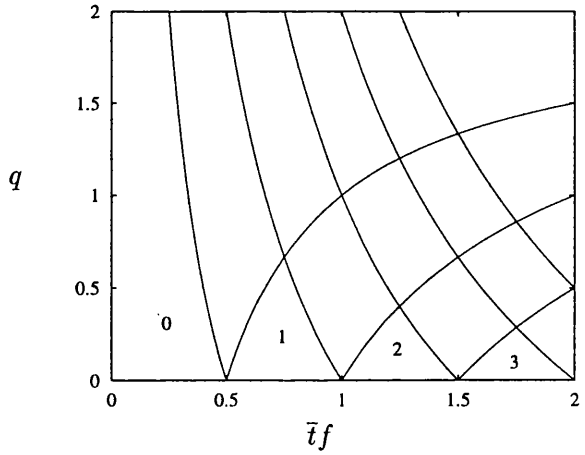


Figure 5.1. Allowed regions $\mathcal{A}_k$ of (q, f) plane for prediction of a sinusoid. Numbers on the graph refer to the index k; unnumbered regions are disallowed. ($\mathcal{A}_0$ is the leftmost strip; $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are the little triangular segments.)

We also examined the effect of increasing the model order to $d = 3$ (Figure 5.3(a,c)) and $d = 4$ (Figure 5.3(b,d)). For these simulations the number of centres was increased to 90 (for $d = 3$) and 120 ($d = 4$). Interestingly these graphs are virtually identical to the previous ones: increasing d seems to have very little effect. This points to the conclusion that even in higher embedding dimensions the prediction function H remains discontinuous when (q, f) is not in an allowed region.

Looking at Figure 5.1 again, it is interesting that the only value of q for which all frequencies are admissible is $q = 0$, which means regular sampling. (Strictly speaking we have not examined the behaviour on the boundaries of the $(\mathcal{A}_k)$.) The case $q = 0$ is easily analysed, for then we obtain an embedding whenever $2\bar{t}f$ is irrational, and not otherwise.

It would be interesting to see how Figure 5.1 would be affected if a periodic signal other than the sinusoid were considered. It seems likely that the answer to

[1] For the simulations $\bar{t}$, the average sample spacing, is 1.

this question would be provided by the number of maxima and minima per cycle. Let us consider perturbations to the sinusoid. If the perturbation did not affect the number of local maxima and minima (so that there was still one of each per cycle) the diagram would only be slightly affected, though possibly the allowed regions would shrink so that they no longer touched. If multiple maxima and minima were introduced, the diagram would become more complicated, and the allowed regions would probably split. For example, if the observation function were $g : \theta \mapsto \cos 2\theta$ then in the ranges shown in Figure 5.1 ($0 < q < 2$, $0 < \bar{t}f < 2$) there would be eight allowed regions.

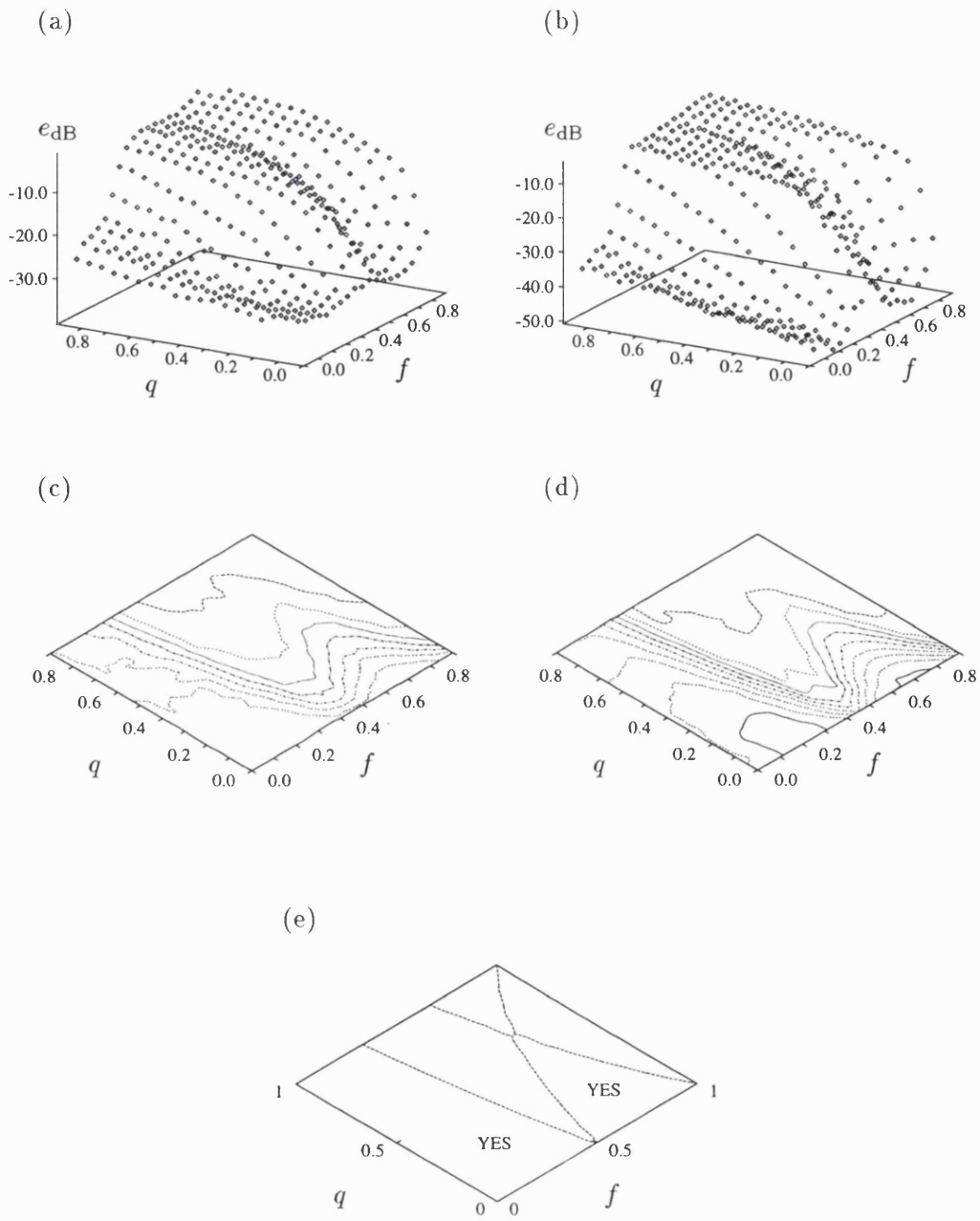


Figure 5.2. Reconstruction errors (embedding dimension 2).
(a,b) Error (dB, top axis) vs. frequency (RH axis) and nonuniformity parameter q (LH axis): (a) 30 centres, (b) 60 centres.
(c,d) Contour plots for surfaces depicted in (a,b).
(e) 'Allowed regions' (marked YES).

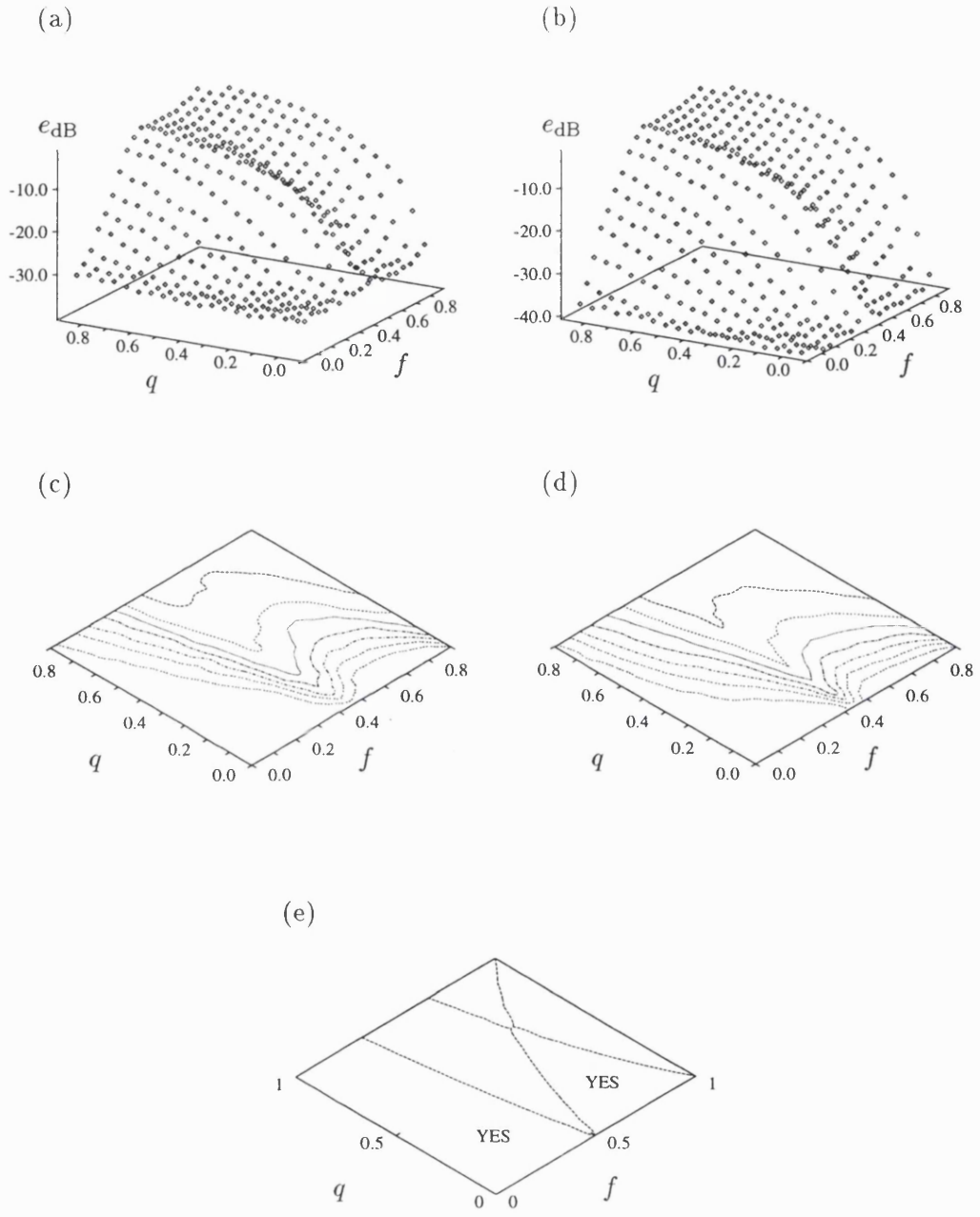


Fig. 5.3. Reconstruction errors (embedding dimension 3,4).
(a,b) Error (dB, top axis) vs. frequency (RH axis) and nonuniformity parameter q (LH axis): (a) $d = 3$, (b) $d = 4$.
(c,d) Contour plots for surfaces depicted in (a,b).
(e) 'Allowed regions' (marked YES).

5.3 Prediction of the Lorenz attractor

The work in the previous section showed that it is possible to predict periodic dynamics when certain conditions are met. In this section we shall consider the prediction of a chaotic dynamical system and show that nonlinear prediction has substantial advantages over linear prediction (classical or generalised) for this type of waveform, which is inherently broad-band and hence not amenable to linear analysis.

The system that we shall choose is the Lorenz attractor [212], governed by the 3-dimensional dynamics

$$\begin{aligned} dZ_1/dt &= 10(Z_2 - Z_1) \\ dZ_2/dt &= -Z_1 Z_3 + 28Z_1 - Z_2 \\ dZ_3/dt &= Z_1 Z_2 - \tfrac{8}{3} Z_3 \,. \end{aligned}$$

Strictly speaking we should not apply Takens' theorem without checking that the underlying manifold is compact. It can be shown using Lyapunov functions [73] that the Lorenz system is Lyapunov stable, i.e. trajectories that start in some closed ball stay within another closed ball (in $\mathbf{R}^3$). Hence the attractor is contained in a compact region, and it can be shown that this is sufficient for the conclusions of Takens' theorem to go through [90, 178].

Suppose that observations are made on the first variable, so that the observed waveform is $x(t)$ defined as

$$x(t) = Z_1(\gamma t). \tag{5.5}$$

As in Chapter 4, where the same signal was used, γ controls the bandwidth of the process: the higher the value of γ, the more broad-band the waveform $x(t)$. We do not wish γ to be too small, for if it were, the signal $x(t)$ would be too slowly-varying and too easy to predict by linear methods, and this would make the comparison between linear and nonlinear techniques rather pointless. In the simulations of this section $\gamma = \frac{1}{7}$. A typical realisation of $x(t)$ is shown in Figure 5.4(a). This was obtained by integrating the Lorenz equations using an adaptive step-size fourth-order Runge-Kutta algorithm[2] and a relative accuracy of 10^{-6}. The sampling could therefore be performed by nominating the next sample point, and integrating the equations as far as that point. The initial conditions, which correspond to a point on or very close to the attractor (as opposed to an arbitrary point in space) were $(Z_i)(0) = (4.1, -0.8, 28.8)$.

The Lorenz attractor has underlying dimensionality $D = 3$ so if the sampling were regular one would by Takens' theorem require a seventh-order nonlinear model $(2D + 1)$ to predict it. However it is often the case with nonlinear prediction that if the observation function is reasonably sensible, that is, not too convoluted, one can often successfully predict the time series using a rather lower embedding dimension than $2D + 1$. In the previous section we saw an example of this with linear prediction for the sinusoid, where $d = 2$ is sufficient even though the Takens' embedding theorem suggests a dimension of 3.

[2] NAG routine D02BAF.

The test signal (5.5) was sampled additive-randomly with intersample spacings taken from the rectangular distribution [0.5,1.5]. (This is the same as the type (b) scheme of §4.4.1, or $q = 1$ in the previous section.) A section of 4000 points was taken. The first 2000 were used for training points (the centres were taken from this section too) and the remaining 2000 were used for verification. As with the test results of §5.2, the prediction errors e_{ms} and e_{dB} were calculated using the verification data.

Results for fourth-order nonlinear prediction ($d = 4$) are shown in Figure 5.4(b), using 180 radial basis centres. The prediction errors are shown as a time series and their energy is 19.0dB below that of the original signal (e_{dB}, in the nomenclature of the previous section, was -19.0dB). We think that this is a reasonable result, but it is not as good as the sort of accuracy one can get for regular sampling. We found that a time series obtained by sampling (5.5) regularly, with intersample spacings all equal to 1, could be predicted with error -32dB (this required 60–70 radial basis centres).

Some further experiments were performed. First, Figure 5.4(c) shows the effect of putting $w_\tau = 0$ rather than the 'automatic' choice (5.3) based on the variance of the intersample spacings (as discussed in §5.1). This means that in performing the prediction the intersample spacings are ignored. The results are seen to be worse ($e_{\text{dB}} = -12.1$dB). So in Figure 5.4(b) the intersample spacings are being put to good use in establishing the prediction function H.

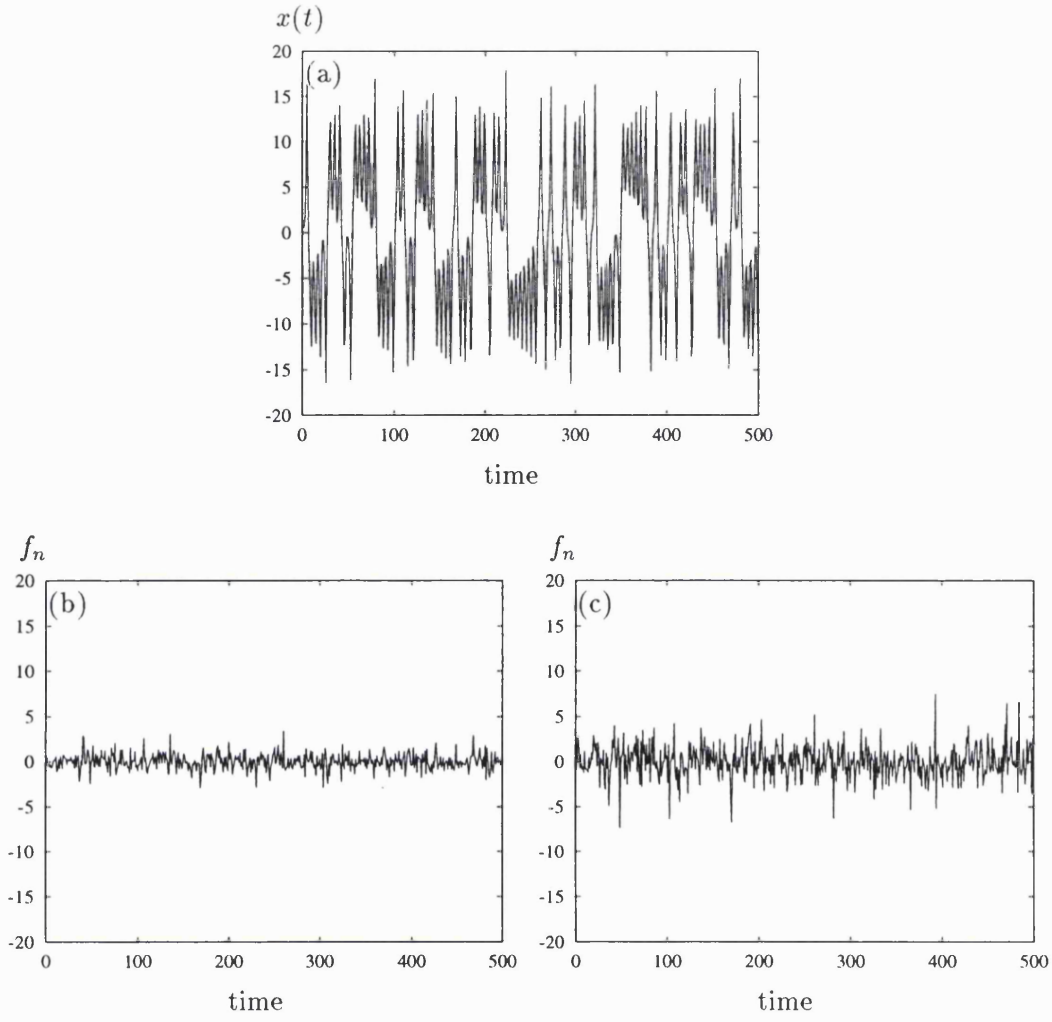


Fig. 5.4. Nonlinear prediction of the Lorenz attractor.
(a) Time series. (b,c) Prediction errors of irregularly sampled record, using 4th-order nonlinear prediction and a 180-centre RBF approximant to H :
(b) automatic selection (5.3) of width parameters $w_{x,\tau}$ (prediction error -19.0dB);
(c) w_τ set to zero, excluding the temporal information (prediction error -12.1dB).

It is also of interest to examine the effect of linear estimation using the generalised prediction error methods of Chapter 4. When these are employed the results are substantially worse; this is to be expected on account of the broad spectrum. Once the 'optimal' model poles $\boldsymbol{\alpha}$ were found the generalised prediction errors $\mathfrak{f}_n(\mathbf{x};\boldsymbol{\alpha})$ were calculated; as they are determined only up to sign, their absolute values are plotted in Figure 5.5(a). The prediction error is -4.8dB.

One might ask at this point, given the obvious failing of the linear technique, whether the underlying signal is fundamentally unpredictable by linear means, or whether our technique for fitting linear models is at fault. That question is easily resolved if one recalls that all a linear method can do is estimate the spectrum, if it is assuming that the statistics are Gaussian (as the least-squares AR approach does); so, if the ARPSD resembles the underlying spectrum, the fitting technique has done as well as could be expected of it. To compare these spectra we have calculated the ARPSD from the model coefficients (Figure 5.5b) and obtained an estimate of the underlying PSD by resampling $x(t)$, regularly with intersample spacing 1, and running a 1024-point windowed FFT (Figure 5.5c). If one assumes the true PSD to be continuous, which seems sensible, then it would resemble a smoothed version of Figure 5.5c—in fact, something very like the AR spectrum. We may conclude that the linear model-fitting technique has indeed chosen the best fourth-order linear model, but that no linear model is a good representation.

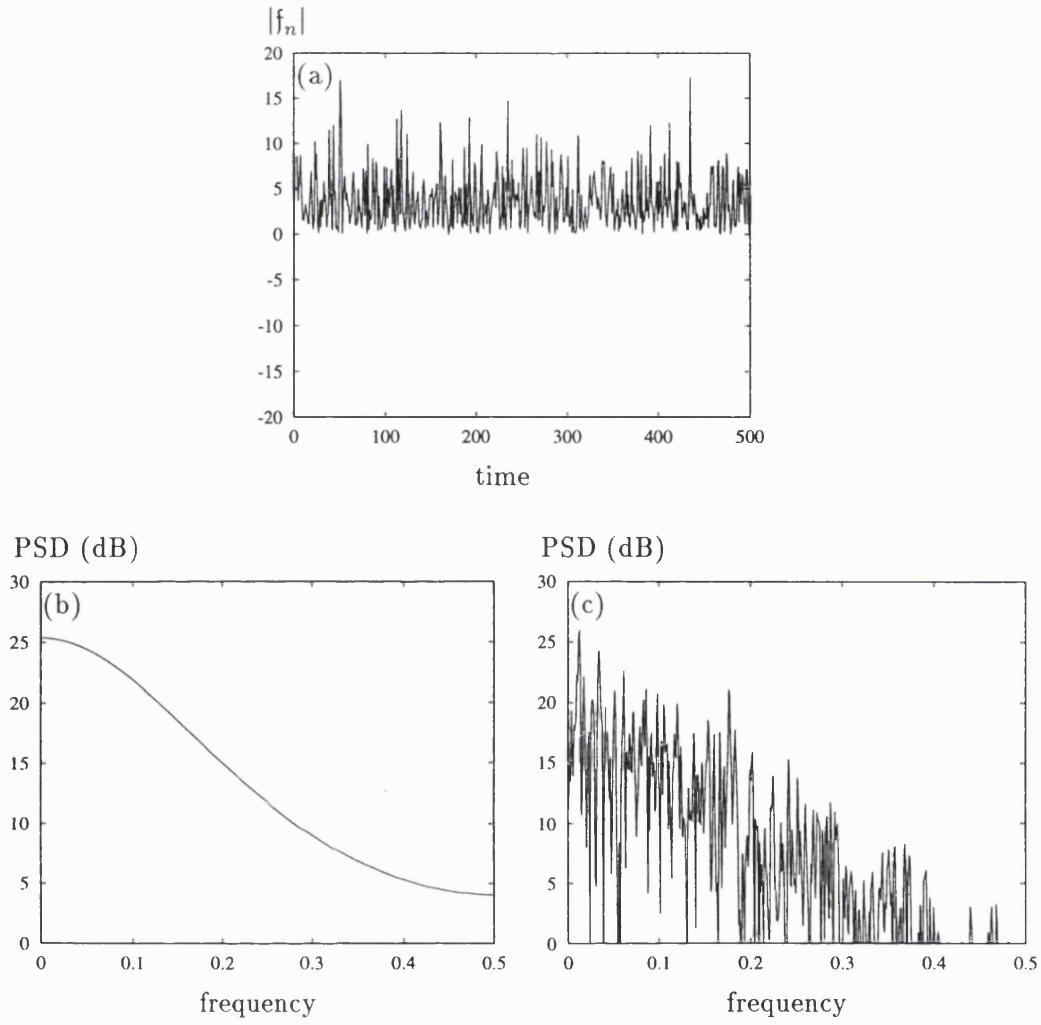


Fig. 5.5 Linear methods applied to the Lorenz attractor.
(a) Generalised prediction errors $|f_n|$ (prediction error -4.8dB).
(b,c) Spectral estimates are in broad agreement:
(b) AR(4) PSD estimate, (c) FFT of regularly sampled record.

5.4 Final remarks

We have found two test cases on which linear and nonlinear prediction give opposite results. For a sinusoid, the concept of *generalised* linear prediction, developed for that purpose, works well, as the results of Chapter 4 showed. Nonlinear prediction does well but not for all frequencies and sampling schemes. There is a simple criterion $(q, f) \in \mathcal{A}$ [eq. (5.4), §5.2.2] that states how irregular the sampling is allowed to be for a given frequency. It bears a passing resemblance to the criteria of Nyquist and of Kadec, for like the former it shows that prediction fails as the frequency gets to half the average sampling rate—from Figure 5.1 we see that the regions $\mathcal{A}_0$ and $\mathcal{A}_1$ touch at the point $(q, \bar{t}f) = (0, 0.5)$—and like the latter it gives a condition on how far the sampling may deviate from uniformity.

This has interesting implications for embedding chaotic dynamics. Chaotic systems have an infinite supply of periodic orbits[3]. Figure 5.1, and the experimental results which corroborate it, shows that, depending on the degree of aperiodicity in the sampling, it is not always possible to embed periodic orbits correctly. We have shown that the Lorenz system can be predicted quite well by nonlinear means when the sampling is irregular (relative error -19dB). But the prediction is not as good as it is when the sampling is regular (-32dB); nor is it as good as nonlinear prediction of periodic dynamics (which, when it works, gives prediction errors of typically -30 to -40 dB). This suggests that in a chaotic system such as the Lorenz system, there is a 'spectrum' of periodic orbits some of which are in the 'allowed regions' of Figure 5.1 and others not. This 'spectrum' is given by a mathematical device known as the zeta-function which gives the density of orbits of period T as a function of T. (It is not the same as the Fourier spectrum, because the orbits are not stable and not observed as sinusoidal functions in the time series.) An interesting question for further research is to see if knowledge of the zeta-function, together with Figure 5.1, directly implies the quality of prediction of irregularly sampled chaotic time series.

We have also been able to tie in the work of Chapter 4. For the Lorenz attractor the linear technique attempts to model it as a stochastic process and gives a good representation of the spectrum, but nonlinear prediction does much better.

There arises the question of whether the two schemes can be fused. Looking at the derivations one sees instantly that the two methods are very different. The linear model makes a very specific assumption about the underlying process and the generalised prediction errors are constructed from the parameters associated with that model. On the other hand the nonlinear model makes very weak assumptions about the underlying process and although this makes it more flexible it also means that much more training data is needed. Certainly there is no notion of a finite set of parameters enabling one to construct generalised prediction errors.

One would therefore have to consider a class of nonlinear dynamical systems that could be parametrised in a simple way, and define a predictor to be a function $\wp : \mathbb{R}^{d+1} \to \mathbb{R}$ depending on the sampling instants and satisfying the condition

$$\wp(x_n, \ldots, x_{n-d}) = 0 \quad \text{for all realisations of the dynamics.}$$

[3]Of course they are unstable, but that is not the issue: they are still in the dynamics, and so have to be embedded.

The set $\mathfrak{P}$ of all such predictors is a linear space (though of course $\wp : \mathbb{R}^d \to \mathbb{R}$ is nonlinear) and one would have to reconsider the question of how to choose an appropriately-normalised predictor $\hat{\wp}$ from $\mathfrak{P}$. The quantity $|\hat{\wp}(x_n, \ldots, x_{n-d})|$ would then be the generalised nonlinear prediction error. The difficulty lies in identifying $\mathfrak{P}$, which appears to involve solving the nonlinear differential equation (not an easy proposition).

Another possible area for research is nonlinear filtering. There is a nice method, due to Broomhead [28, 29], for constructing 'nonlinear inverses to linear filters' when the sampling is regular. This is how they work. Recall that, in the notation we are using, Φ is the diffeomorphism between the trajectory in $\mathbf{R}^d$ described by the sequence of delay-vectors and the trajectory followed by the system in its phase space. It turns out that there is still such a diffeomorphism even if the observed time series data are passed through a linear FIR filter. Let Γ denote the FIR filter, which acts in the obvious way on $\mathbf{E}$ by

$$\Gamma : (x_{n-i})_{i=1}^{d} \mapsto \left(x_{n-i} + \sum_{j=1}^{p} \gamma_j x_{n-i-j} \right)_{i=1}^{d}.$$

If we write Φ^Γ for the Takens embedding associated with the time series filtered by Γ, then the following diagram can be constructed:

$$\begin{array}{ccc} \Gamma\mathbf{E} & \xrightarrow{K^\Gamma} & \Gamma\mathbf{E} \\ {\scriptstyle \Phi^\Gamma}\uparrow & & {\scriptstyle \Phi^\Gamma}\uparrow \\ \mathcal{M} & \xrightarrow{\phi} & \mathcal{M} \\ {\scriptstyle \Phi}\downarrow & & {\scriptstyle \Phi}\downarrow \\ \mathbf{E} & \xrightarrow{K} & \mathbf{E} \end{array}$$

We are given a sum $\{z_n\}$ of two signals, $\{x_n\}$ dynamical, and $\{u_n\}$ lying in some known spectral band so that a filter Γ can be found to remove it (or at least substantially reduce its magnitude). We wish to recover $\{u_n\}$ from $\{z_n\}$. Note that, by linearity of Γ,

$$\Gamma(\underline{\mathrm{x}}^n) = \Gamma(\underline{\mathrm{x}}^n) + \Gamma(\underline{\mathrm{u}}^n) = \Gamma(\underline{\mathrm{x}}^n + \underline{\mathrm{u}}^n) = \Gamma(\underline{\mathrm{z}}^n) \in \Gamma\mathbf{E}.$$

The map K takes $\underline{\mathrm{x}}^n$ to $\underline{\mathrm{x}}^{n+1}$; as before, write H for the component of K that takes $\underline{\mathrm{x}}^n$ to x_n. Now all we have to do now is diagram-chase:

$$u_n = z_n - x_n = z_n - H(\underline{\mathrm{x}}^n) = z_n - H\Phi(\Phi^\Gamma)^{-1}(\Gamma\underline{\mathrm{z}}^n)$$

so that

$$\Phi(\Phi^\Gamma)^{-1} : \Gamma\mathbf{E} \to \mathbf{E}$$

is a 'nonlinear inverse' to the linear filter Γ. One needs to know Γ and have a suitably long stretch of data from the dynamics; in advance of somebody giving us the time series $\{z_n\}$ we have representations for $\mathbf{E}$ and (on applying Γ) $\Gamma\mathbf{E}$ and can construct

the map $\Phi(\Phi^\Gamma)^{-1}$. Broomhead *et al.* [28, 29] performed the construction using a radial basis function fit.

This is a subtle approach and (not surprisingly) the key step is to show that Φ^Γ is injective. An elementary argument suggests that chaotic signals have broad spectra and so spread their information across the spectrum; applying a linear filter cannot remove all the information, and so a Takens embedding should still exist. Obviously this is not true for linear processes: for example applying an FIR filter to a bunch of sinusoids could cause some of the tones to be removed completely, and then the filter could not be inverted.

Having studied prediction techniques for irregularly sampled nonlinear data, we are able to construct H or, equivalently, K. Also the techniques of Chapter 4 enable the construction of narrow band-stop filters. Hence the nonlinear inverse filter should be constructible. The only source of concern is that the filters we have developed are in general IIR and acausal. IIR filters can generate new dynamics, depending on how quickly the impulse response decays [201, 203]; if they do generate new dynamics, the dimensionality of the time series is increased by filtering. This might require the theory to be extended. Presumably in the above diagram the top branches would have to be altered, for the inclusion of the filter dynamics in $\Gamma\mathbf{E}$ would mean that Φ^Γ was no longer onto $\Gamma\mathbf{E}$, but rather onto a nonlinear subspace of it.

Chapter 6

Conclusions

This thesis has brought techniques from the theories of integral transforms, complex analysis, local fields, linear stochastic processes, and dynamical systems, to bear upon the problem of analysing signals sampled at irregular time instants. We have deliberately reserved specific comments on the various methods for their respective chapters: the purpose of this chapter is to make some observations on the state of the theory as we have left it.

The aim of this thesis has been to answer the three questions listed in the Introduction, which we reproduce here for convenience:

Q1 How do irregular samples of an underlying waveform relate to that waveform?

Q2 Given an irregularly sampled signal, how should we process it?

Q3 How should we generalise common signal processing techniques such as filtering, convolution, model fitting, etc., to cope with irregularly sampled data?

The main theme has been the analysis and processing of irregularly sampled data without recourse to explicit signal reconstruction.

From the point of view of practical signal processing the key developments have been the linear prediction and filtering schemes of Chapter 4. This has produced algorithms for clutter modelling, clutter removal and noise reduction. In addition to being used on their own, these can also be combined with elementary spectral analysis to solve the problem of signal identification.

The linear work has raised many questions of its own and notwithstanding a couple of awkward corners in the theory it seems that exploitation of the relationship between continuous- and discrete-time theory has been a fruitful line of investigation. Of course the notion of time-varying prediction was always likely to be right approach, but exactly how to make the predictors vary is a rather difficult problem. What was not so obvious is that it is not necessarily a good idea to predict an observation from a finite number of previous ones: the prediction coefficients only have to be capable of generating a bilinear function (of themselves and the data) that gives an indication of how well a putative model accords with the data that one is trying to fit. The construction of such an error energy function, playing the role of a

log-likelihood function, then allows the construction of a Wiener filter in a way that is almost disappointingly straightforward. We suggest that it is now time for this filtering technique to take its place in the signal processing toolbox, as a method for separating a narrow-band signal (from a combination of it and something else) when the sampling is irregular; and it should fit well into modern technology that looks uses irregular sampling 'in anger' [20, 21, 14, 120].

Chapter 3 discussed the theory of sampled signals. In §3.1 we showed that a more general type of integral transform is sufficient to derive a theory of sampling, convolution products and Volterra operators, provided that the transform kernel satisfies a 'completeness' condition (namely that the kernels $K(u, \ell_n)$ were an orthogonal basis for $L^2(I)$, where I is the set on which the transform of the signal is supported). Hence we have shown that periodicity of the sampling is not required for all this to go through, though it must be said that the examples are easier when the sampling is regular.

In §3.2 we contributed to the theory of contour integration, as applied to the derivation of error bounds for sampling series. This method is likely to prove useful partly as a tool in approximation theory and partly because of its connection with the theory of canonical products. There is plenty of scope for ingenuity here, for in principle any function with poles can be used as an integration kernel. However it is not quite as simple as that, because one needs to ensure that (i) a workable upper bound for the integrand can be found, and (ii) the truncation error bound tends to 0 as the number of samples tends to ∞ (otherwise, the resulting sampling series is of little use).

Methods from local field theory in §3.3 gave a new viewpoint for analysing signals that are generated by exponentials and polynomials. It is interesting that they are able to solve these problems so neatly, and one is left asking why exponentials managed to creep back into a subject that did not seem to require them. The most convenient answer is that one is examining functions defined on $\mathbb{Z}$, and that a key point is to find a signal vanishing at each integer—whereupon the sinusoid enters. An open question is whether the class of functions, for which those theorems were derived, can be expanded.

Finally the young subject of nonlinear signal processing entered in Chapter 5. Nonlinear methods have produced prediction schemes for data that are of dynamical origin. The approach is concerned with the fundamental result that underlying dynamics confer a property of predictability on the sampled data. In nonlinear filtering—a difficult problem even in the regular case because a nonlinear filter does not handle in a simple way an additive combination of signals—there seem to be interesting possibilities for further research. Described in §5.4 is Broomhead's method [28, 29] for separating a dynamical signal (we know the dynamics) from an additive combination of it and an arbitrary signal that can be cancelled with a known linear FIR filter. Using the dynamics one can construct a 'nonlinear inverse' which recovers the dynamical signal from delay-vectors that have been linearly filtered. One takes the time series, filters it using the linear FIR filter, and then applies the 'nonlinear inverse' to recover the dynamical signal. When the sampling is irregular we can construct nonlinear predictors and linear filters using the techniques in Chapters 4

and 5, so that should allow nonlinear filters to be constructed for aperiodic data.

When the sampling is periodic, nonlinear prediction can be regarded as subsuming linear prediction (albeit with significantly increased opacity and computational expense). When the sampling is not periodic, the complications that are thereby introduced make the subjects distinct (at least, they do at present). This is because the nonlinear prediction is attempting to predict an observation from previous values and their associated intersample spacings, whereas the linear methods employ the method of *generalised* prediction referred to above. So in our investigations we have found data that are well-handled by generalised linear prediction but not by the nonlinear method, and *vice versa*. It is apparent, then, that the different techniques for analysis discussed in this thesis are most suited to analysing 'their own types of signal', and that leaves future researchers with a rich supply of avenues to investigate.

In conclusion we may say that, although there is plenty still to be done on all three questions, we have narrowed the gap between irregular sampling as an area for purely academic research and irregular sampling as a powerful and practical tool.

Appendix A

Spectrum of nonuniform samples

This Appendix gives derivations for the spectrum of samples, $P^s(f)$, of a process, as a function of its underlying spectrum $P(f)$. The case of jittered and additive-random sampling are considered. These issues were first addressed by Shapiro & Silverman [189] who preferred to work with the autocorrelation. Although they hint at the result

$$P^s = P * \mathrm{SC}$$

they do not explicitly state the form of SC for jittered sampling, and for additive-random sampling they incorrectly derive $P^s(f)$ from the sample autocorrelation ([189], p.236).

The following notation will be used.

- $x(t)$, continuous-time stationary process
- $R(\tau)$, underlying autocorrelation. The conventional definition for this is $\mathcal{E}^x x(t)^* x(t+\tau)$ which is independent of t provided that x is wide-sense stationary. In that case it can be replaced by a time-average to get $\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)^* x(t+\tau) dt$.
- $P(f)$, underlying power spectrum :

$$P(f) = \lim_{T\to\infty} \frac{1}{2T} \left| \int_{-T}^{T} x(t) e^{-2\pi i f t} dt \right|^2$$

- $P^s(f)$, observed power spectrum :

$$P^s(f) = \lim_{N\to\infty} \frac{1}{2N} \left| \sum_{n=-N}^{N-1} x_n e^{-2\pi i f t_n} \right|^2$$

- p_m, pdf of intersample spacing $t_{n+m} - t_n$.

- $\tilde{}$ denotes the Fourier transform : $\tilde{p}(f) \equiv \int_0^\infty p(t)e^{-2\pi i f t}dt$.
- $\bar{t}$, mean intersample spacing

For additive-random sampling :

- p, pdf of intersample spacing $t_{n+1} - t_n$. Note that $p(t) = 0$ for $t < 0$.
- Here p_m is the m-fold self-convolution of p, and so the Fourier transform of $p_m(t)$ is $\tilde{p}(f)^m$.

For jittered sampling :

- q, the pdf of the jittering. Thus $t_n = (n + s_n)\bar{t}$ where s_n are i.i.d. observations having this pdf.
- Q, the pdf of $s_{n+m} - s_n$. For $m \neq 0$ we have

$$\begin{aligned} Q(s) &= \int q(s')q(s'+s)\,ds' \\ \tilde{Q}(f) &= |\tilde{q}(f)|^2 \end{aligned}$$

- Here $p_m(t)$ is $\delta(t)$ when $m = 0$, and $\bar{t}^{-1}Q(t/\bar{t} - m)$ for all other m.

A.1 Additive-random sampling

We have

$$P^s(f) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{|m|\leq N} \sum_{n=1}^{N} x_n^* x_{n+m} e^{-2\pi i f(t_{n+m}-t_n)}$$

We now split the m-summation into three parts : the first has $m = 0$, the second contains the terms with $m > 0$, and the third has $m < 0$. Clearly the third part is the complex conjugate of the second, so we shall write it as 'c.c.' from now on. Then, replacing the n-summation with an integral w.r.t. the distribution function of $(t_{n+m} - t_n)$ we have

$$P^s(f) = R(0) + \left(\sum_{m=1}^{\infty} \int_0^\infty R(\tau)e^{-2\pi i f\tau} p_m(\tau)d\tau + \text{c.c.} \right)$$

Using the Wiener-Khinchine theorem,

$$R(\tau) = \int_{-\infty}^{\infty} P(f')e^{2\pi i f'\tau} df' ,$$

we have

$$P^s(f) = \int_{-\infty}^{\infty} P(f')df' + \left(\sum_{m=1}^{\infty} \int_0^\infty \int_{-\infty}^{\infty} P(f')e^{2\pi i(f'-f)\tau} p_m(\tau)\,df'\,d\tau + \text{c.c.} \right)$$

or

$$\begin{aligned} P^s &= P * \mathrm{SC} \\ \mathrm{SC}(f) &= 1 + \left(\sum_{m=1}^{\infty} \int_0^{\infty} e^{-2\pi i f \tau} p_m(\tau)\, d\tau + \text{c.c.} \right) \\ &= 1 + \left(\sum_{m=1}^{\infty} \tilde{p}(f)^m + \text{c.c.} \right) \\ &= \mathrm{Re}\left(\frac{1+\tilde{p}(f)}{1-\tilde{p}(f)} \right) \quad \textbf{except when } \tilde{p}(f) = 1 \end{aligned} \tag{A.1}$$

The question of what happens when $\tilde{p}(f) = 1$ is a delicate one; we contend that at such points, $\mathrm{SC}(f)$ has a delta-function singularity. Let the frequencies for which $\tilde{p}(f) = 1$ be called 'singular frequencies' and denoted f_d. Note that 0 is always a singular frequency, and that the singular frequencies must be isolated ($\tilde{p}$ is analytic). Let us consider the behaviour of $\mathrm{SC}(f)$ at the origin. Expanding $\tilde{p}(f)$ as a Taylor series (valid as $\tilde{p}$ is analytic) in the vicinity of $f = 0$,

$$\tilde{p}(f) = 1 + \tilde{p}'(0) f + \cdots,$$

we have in the limit $\varepsilon \to 0^+$,

$$\begin{aligned} \int_{-\varepsilon}^{+\varepsilon} \mathrm{SC}(f)\, df &= \mathrm{Re} \int_{\Gamma} \left(\frac{2 + \tilde{p}'(0) f + \cdots}{-\tilde{p}'(0) f + \cdots} \right) df \\ &= \mathrm{Re} \left\{ \pi \mathrm{i}\, \mathrm{Res}\left(\frac{-2}{\tilde{p}'(0) f}, f = 0 \right) \right\} \\ &= \mathrm{Re}\, \frac{-2\pi \mathrm{i}}{\tilde{p}'(0)} \end{aligned}$$

in which the contour Γ runs from $-\varepsilon$ to $+\varepsilon$ in the *lower* half of the complex plane. This is because $\mathrm{SC}(f)$ is defined as an infinite series (by A.1) and so for validity we must keep $|\tilde{p}(\Gamma)|$ inside the unit circle; given that $|\tilde{p}(f)| < 1$ in the lower (not the upper) half-plane, it is in that half-plane that we must put the contour. Now

$$\tilde{p}'(0) = -2\pi \mathrm{i} \int_0^{\infty} t p(t)\, dt = -2\pi \mathrm{i} \bar{t}, \qquad \bar{t} = \text{mean sample spacing}$$

and so $\mathrm{SC}(0)$ is a delta-function of strength $1/\bar{t}$. By an identical argument, a singular frequency f_d generates a delta-function in $\mathrm{SC}(f)$ of strength $h_d := \mathrm{Re}\,[-2\pi \mathrm{i}/\tilde{p}'(f_d)]$ at $f = f_d$. Thus

$$\mathrm{SC}(f) = \delta(\bar{t} f) + \mathrm{Cont}\, \mathrm{Re} \left(\frac{1+\tilde{p}(f)}{1-\tilde{p}(f)} \right) + \sum_{f_d \neq 0} h_d\, \delta(f - f_d) \tag{A.2}$$

where 'Cont' signifies that the $\mathrm{Re}(\cdot)$ term is taken as continuous at the (isolated) singular frequencies.

We now consider the mean of the function SC. By the Riemann-Lebesgue Lemma we have that whenever p is a measurable pdf, $\tilde{p}(f) \to 0$ as $f \to \pm\infty$. Hence the average of $\mathrm{Cont}\, \mathrm{Re} \left(\frac{1+\tilde{p}(f)}{1-\tilde{p}(f)} \right)$ is 1. For a measurable pdf there is no delta-function train, so the average of $SC(f)$ as given by (A.2) over $f \in \mathbb{R}$ is 1.

A.2 Jittered sampling

We have

$$P^s(f) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{|m|\le N} \sum_{n=1}^{N} x_n^* x_{n+m} e^{-2\pi i f(t_{n+m}-t_n)}$$

We bring the $m = 0$ term out separately, where as before it generates $R(0)$. Then, replacing the n-summation with an integral w.r.t. the distribution function of $(t_{n+m} - t_n)$ we have

$$\begin{aligned} P^s(f) &= R(0) + \sum_{m\neq 0} \int R(\tau) e^{-2\pi i f\tau} Q(\tau/\bar{t} - m)\, d\tau/\bar{t} \\ &= R(0) + \sum_{m\neq 0} \int R((m+\sigma)\bar{t}) e^{-2\pi i f\bar{t}(m+\sigma)} Q(\sigma)\, d\sigma. \end{aligned}$$

Using the Wiener-Khinchine theorem as before,

$$P^s(f) = \int_{-\infty}^{\infty} P(f')df' + \sum_{m\neq 0} \int \int_{-\infty}^{\infty} P(f') e^{2\pi i(f'-f)\bar{t}(m+\sigma)} Q(\sigma)\, df'\, d\sigma$$

or $P^s = P * \mathrm{SC}$ with

$$\begin{aligned} \mathrm{SC}(f) &= 1 + \sum_{m\neq 0} \int e^{-2\pi i f\bar{t}(m+\sigma)} Q(\sigma)\, d\sigma \\ &= 1 + |\tilde{q}(\bar{t}f)|^2 \sum_{m\neq 0} e^{-2\pi i m f\bar{t}} \\ &= 1 - |\tilde{q}(\bar{t}f)|^2 + |\tilde{q}(\bar{t}f)|^2 \sum_{n\in\mathbb{Z}} \delta(\bar{t}f - n). \end{aligned} \tag{A.3}$$

Notice that the delta-function train characteristic of regular sampling is still present, so that aliases are not suppressed as effectively as they are when the sampling is additive-random. An interesting case is when q is the rectangular distribution on $[-\frac{1}{2}, \frac{1}{2}]$, which corresponds to uniform '100%' jitter (i.e. successive samples can actually touch). Then $\tilde{q}(u) = \operatorname{sinc} u$ which vanishes at all nonzero integers, and one is left with a delta-function at the origin, $\delta(\bar{t}f)$, plus a continuum given by $1 - \operatorname{sinc}^2 \bar{t}f$.

Concerning the average of SC over $f \in \mathbf{R}$, we can say that for a measurable pdf p, $\tilde{q}(f) \to 0$ as $f \to \pm\infty$, and so only the first term in (A.3) contributes. So the average is 1.

Appendix B

Introduction to local field theory

The reader is referred to Cassels [44] for a concise treatment or Schikhof [181] for a more extensive one. Cassels concentrates on the algebraic aspects, Schikhof on those related to p-adic Analysis. Elementary results on groups, rings and fields are dealt with by, for example, Stewart [204], while Baker's book [9] discusses the relevant bits of number theory.

B.1 Valuations

Most common properties of the standard ('archimedean') valuation on $\mathbb{Q}$ are corollaries of the following three :

V1 Positivity : $|x| \geq 0$ with equality iff $x = 0$

V2 Multiplicativity : $|xy| = |x||y|$

V3 Triangle inequality : $|x + y| \leq |x| + |y|$

The standard valuation is not the only one to do so; there are a family of others, known as the p-adic valuations, which satisfy the above three conditions, and also this one,

V4 Ultrametric inequality : $|x + y| \leq \max(|x|, |y|)$

which is stronger than V3. Such valuations are said to be *non-archimedean.*

The ultrametric inequality gives rise to some interesting properties. Let K be a field with nonarchimedean valuation $|\cdot|$. Then we define

- $\mathfrak{o}$, the elements in K satisfying $|x| \leq 1$. This is closed under multiplication *and under addition*, so it is a ring, called the *valuation ring.* It is also a *local ring* as it has a unique maximal ideal.

- $\mathfrak{m}$, the elements in K satisfying $|x| < 1$. It is the maximal ideal of $\mathfrak{o}$.

- $\mathfrak{u}$, the 'units' of K, satisfying $|x| = 1$.
- $k = \mathfrak{o}/\mathfrak{m}$, the *residue class field* of K. The mapping from $\mathfrak{o}$ to $\mathfrak{o}/\mathfrak{m}$ will be called λ. It is easy to see that $x \in \mathfrak{m} \Leftrightarrow \lambda x = 0_k$ and $x \in \mathfrak{u} \Leftrightarrow \lambda x \in k^\times$ (the nonzero elements of k).
- $\mathrm{Pos}(K)$, the elements of K satisfying $|x-1| < 1$. Then $x \in \mathrm{Pos}(K) \Leftrightarrow \lambda x = 1_k$. From this we see that $\mathrm{Pos}(K)$ is a group under multiplication.
- V, the set of $\{|x| : x \in K^\times\}$, clearly a subgroup of $\mathbf{R}^+$, and called the *value group.*

The following is a characterisation of $\mathrm{Pos}(K)$, which will be useful later :

Lemma 1. Let K be a field with a nonarchimedean valuation $|\cdot|$ and suppose that the residue class field k has p elements. If $\alpha \in K$ obeys $|\alpha| = 1$ then $\alpha^{p-1} \in \mathrm{Pos}(K)$.

For the proof, note that $\lambda\alpha \in k^\times$, so $\lambda(\alpha^{p-1}) = (\lambda\alpha)^{p-1} = 1_k$ (Fermat's little theorem). So $\alpha^{p-1} \in \mathrm{Pos}(K)$. □

Corollary 2. K, k, $|\cdot|$ as above. If $|\alpha| = 1$ and $(p-1) \mid m$ then $\alpha^m \in \mathrm{Pos}(K)$. □

That is the end of the algebraic content we need; we now state a remarkable analytical result.

Theorem 3. A series converges ultrametrically if (!) and only if its terms tend to zero.

By convergent, we mean of course that the sequence of partial sums tends to a limit. Write $s_n = \sum_{i=1}^{n} a_i$. Let $\varepsilon > 0$, and choose N s.t. $n > N \Rightarrow |a_n| < \varepsilon$. Then

$$m > n > N \Rightarrow |s_m - s_n| = |a_{n+1} + \cdots + a_m| \leq \max\{|a_{n+1}|, \ldots, |a_m|\} < \varepsilon;$$

now use the General Principle of Convergence. □

We now introduce the p-adic numbers.

Definition 4. For any prime p the p-adic valuation $|\cdot|_p$ of a rational number $rs^{-1}p^\nu$, with r and s coprime to p, is $p^{-\nu}$.

It is easy to verify that V1–V4 are obeyed. The chief difficulty with such valuations is psychological : numbers are p-adically small if they contain large numbers of p's in the numerator, e.g.

$$(p = 2) \qquad 1, 2, 4, 8, \ldots \to 0.$$

Definition 5. The *completion* of a field with respect to the valuation $|\cdot|$ is defined as the set of all limits of convergent (w.r.t. $|\cdot|$) sequences of elements in that field. For example the completion of $\mathbf{Q}$ w.r.t. the standard valuation is $\mathbf{R}$. Essentially the completion process 'fills in the gaps'.

Definition 6. The p-adic field $\mathbf{Q}_p$ is the completion of $\mathbf{Q}$ w.r.t. $|\cdot|_p$. Its valuation ring is called $\mathbf{Z}_p$ and can be identified as the set of series

$$a_0 + a_1 p + a_2 p^2 + \cdots \qquad (0 \leq a_i \leq p-1)$$

The maximal ideal of $\mathbb{Z}_p$ is the set of series with $a_0 = 0$. The residue class field is therefore $\mathbb{F}_p$, the finite field with p elements. The value group of $\mathbb{Q}_p$ is the set $\{p^{-r} : r \in \mathbb{Z}\}$, which incidentally is the same as the value group of $\mathbb{Q}$. From this we see that $\alpha \in \mathrm{Pos}(\mathbb{Q}_p) \Leftrightarrow |\alpha - 1| \leq p^{-1}$, as there is nothing in the value group between 1 and p^{-1}.

For example $-\frac{1}{2} \in \mathbb{Z}_3$, for two reasons : first, its 3-adic value is clearly 1, and secondly, we have a series representation of it,

$$(p = 3) \qquad 1 + 1.3 + 1.3^2 + 1.3^3 + \cdots = \frac{1}{1-3} = -\tfrac{1}{2}.$$

Theorem 7 (Embedding). Let K be a finitely-generated extension of $\mathbb{Q}$. Then K can be embedded in $\mathbb{Q}_p$ for infinitely many choices of p. [In this context an embedding is an injective field homomorphism.]

For the proof, see [44]. This is an important result because it means that one has the benefits of working modulo p but also that K sits inside $\mathbb{Q}_p$ in a natural way.

B.2 Ultrametric functions

In this subsection we quote an important result on the exponential, or power, function.

Definition 8. Let $\mathcal{P}$ denote the set of convergent power series on $\mathfrak{o}$. Then $\mathcal{P}$ is a ring. Also let $\mathcal{W}_p$ be the set of functions from $\mathbb{Z}$ to $\mathbb{Q}_p$ that are p-adically reconstructible, i.e.

$$\{x_n\} \in \mathcal{W}_p \iff \exists g \in \mathcal{P} \text{ s.t. } g(n) = x_n \text{ for all } n \in \mathbb{Z}.$$

Lemma 9. For any positive integer n

$$\left|\frac{1}{n!}\right|_p \leq p^{\frac{n-1}{p-1}}.$$

Proof. A well-known result of elementary number theory asserts that the number of factors of p in $n!$ is

$$\mathrm{ord}_p(n!) = \left\lfloor \frac{n}{p} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \left\lfloor \frac{n}{p^3} \right\rfloor + \cdots$$

Write $d = \lfloor \log_p(n) \rfloor$. Then

$$\mathrm{ord}_p(n!) \leq \frac{n}{p} + \frac{n}{p^2} + \cdots + \frac{n}{p^d} = \frac{n}{p-1}\left(1 - \frac{1}{p^d}\right) \leq \frac{n-1}{p-1}$$

and now the result follows, because $|m|_p \equiv p^{-\mathrm{ord}_p(m)}$. □

Lemma 10. The function $n \mapsto \alpha^n$ is in $\mathcal{W}_p$ if

$$|\alpha - 1|_p < p^{\frac{1}{1-p}}.$$

Proof. The objective is to write $\alpha^\nu = (1+(\alpha-1))^\nu$ and use the binomial expansion of $(1+x)^\nu$. Define $(\nu)_j \equiv \nu(\nu-1)\cdots(\nu-j+1)$ (for $j \geq 1$) and $(\nu)_0 \equiv 1$. (This notation is not standard.) Consider the power series

$$g(\nu) = \sum_{j=0}^{\infty} (\nu)_j \frac{1}{j!}(\alpha-1)^j.$$

If $|\nu|_p \leq 1$ then $|(\nu)_j|_p \leq 1$ and so the jth term in the expansion is bounded by

$$1 \times p^{\frac{j-1}{p-1}} \times |\alpha-1|_p^j \to 0 \text{ as } j \to \infty$$

and, we recall, this is sufficient for the series to converge. So g is defined on $\mathfrak{o}$ as a convergent power series. And, for integer n,

$$g(n) = \sum_{j=0}^{\infty} \frac{(n)_j}{j!}(\alpha-1)^j = \sum_{j=0}^{n} \binom{n}{j}(\alpha-1)^j = (1+(\alpha-1))^n = \alpha^n$$

as required. □

Theorem 11 (Strassmann). Let $g(t)$ be a power series convergent for $t \in \mathfrak{o}$. If g has infinitely many zeros in $\mathfrak{o}$ then g is identically zero. □

Note the connection with complex variable theory—if a function is analytic on the *closed*[1] unit disc and has infinitely many zeros therein, it must be identically zero (the Identity Theorem). But in Ultrametric Analysis the so-called unit disc $\mathfrak{o}$ is very different from what we are used to in Complex Analysis, and the results are often surprising :

Corollary 12. If g is a convergent power series on $\mathfrak{o}$ and g is periodic, then it is constant. For $g(t) - g(0)$ is zero in $\mathfrak{o}$ infinitely many times. □

This really is rather remarkable, and caused by the fact that $\mathfrak{o}$ contains the integers; the fact that 'all the integers are no larger than 1' is a conceptually strange one, but the reader might have gained the impression that Ultrametric Analysis is somewhat more elegant than its Real counterpart, and for those classical functions to which it can be applied (such as polynomials and exponentials) it can be very powerful.

[1] $\sin(1/(1-t))$ is analytic on the open disc and has infinitely many zeros therein, but their limit point is outside.

Bibliography

[1] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring. The analysis of observed chaotic data in physical systems. *Rev. Mod. Phys.*, 65:1331–1392, 1993.

[2] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions.* Dover, New York, 1964.

[3] T. H. Ahberg, E. N. Nilson, and T. L. Walsh. *The theory of splines and their applications.* Academic, New York, 1964.

[4] M. H. Annaby. Kramer-type sampling theorems associated with Fredholm integral equations. *Meth. Appl. Anal.*, 2:76–91, 1995.

[5] M. H. Annaby. One and multidimensional sampling theorems associated with Dirichlet problems. Technical report, Lehrstuhl A für Mathematik, Aachen, 1997.

[6] M. H. Annaby. A sampling theorem for discrete transforms. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 339–342, 1997.

[7] K. Araki. Sampling and recovery formula in finite Hankel transform based on the minimum norm principle. *Trans. Inst. Electron. Commun. Eng. Jpn.*, J68A:643–649, 1985.

[8] S. A. Avdonin. On the question of Riesz bases of exponential functions in L^2. *Vestnik Leningrad Univ. Math.*, 7:5–12, 1979. This is the reference to the translation. The original paper was written in 1974.

[9] A. Baker. *A concise introduction to the theory of numbers.* CUP, 1984.

[10] R. Banning and W. L. de Koning. Spectral analysis of laser Doppler anemometry velocity measurements in fluid flows. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 121–126, 1997.

[11] M. S. Bartlett. On theoretical specification of sampling properties of autocorrelated time series. *J. Roy. Stat. Soc. B*, 8:27–41, 1946.

[12] J. Belcher, J. S. Hampton, and G. Tunnicliffe Wilson. Parametrization of continuous time autoregressive models for irregularly observed time series data. *J. Roy. Stat. Soc. B*, 56:141–155, 1994.

[13] J. J. Benedetto and G. Zimmermann. Sampling multipliers and the Poisson summation formula. *J. Fourier Analysis and Applications*, 3:505–523, 1997.

[14] C. Bérenguer. *Estimation fréquentielle et filtrage de signaux non uniformément échantillonnés: Application à l'estimation radar Doppler non ambiguë.* PhD thesis, University of Nice-Sophia Antipolis, France, 1994.

[15] F. J. Beutler. Sampling theorems and bases in a Hilbert space. *Information and Control*, 4:97–117, 1961.

[16] F. J. Beutler. Error-free recovery of signals from irregularly spaced samples. *SIAM Review*, 8:328–335, 1966.

[17] F. J. Beutler. Alias-free randomly timed sampling of stochastic processes. *IEEE Trans. Info. Theory*, 16:147–152, 1970.

[18] F. J. Beutler and O. A. Z. Leneman. Random sampling of random processes: stationary point processes. *Information and Control*, 9:325–344, 1966.

[19] F. J. Beutler and O. A. Z. Leneman. The theory of stationary-point processes. *Acta Math.*, 116:159–197, 1966.

[20] I. Bilinskis. Deliberately pseudorandomized DSP—advantages and limitations. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 201–207, 1995.

[21] I. Bilinskis. Digital alias-free signal processing and a test-bed for this technology [plus other papers]. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97)*, pages 349–383, Aveiro, Portugal, 1997. These papers are the combined output of the DASP groups in Riga and London.

[22] I. Bilinskis and A. Mikelsons. Application of randomized or irregular sampling as an anti-aliasing technique. In L. Torres, E. Masgrau, and M. A. Lagunas, editors, *Signal Processing V: Theories & Applications*, pages 505–508. Elsevier, 1990.

[23] I. Bilinskis and A. Mikelsons. *Randomized Signal Processing.* Prentice Hall, Cambridge, 1992.

[24] D. R. Brillinger. The spectral analysis of stationary interval functions. *Proc. Sixth Berkeley Symp. Prob. Statist.*, pages 483–513, 1972.

[25] D. R. Brillinger. Statistical inference for irregularly observed processes. In E. Parzen, editor, *Time series analysis of irregularly observed data*, pages 38–57. Springer, 1984.

[26] P. J. Brockwell and R. A. Davies. *Time series: theory and methods.* Springer-Verlag, New York, 1991.

[27] P. J. Brockwell, R. J. Hyndman, and G. K. Grunwald. Continuous-time threshold autoregressive models. *Statistica Sinica*, 1:401–410, 1991.

[28] D. S. Broomhead and J. P. Huke. Signals in chaos: A method for the cancellation of deterministic noise from discrete signals. *Physica D*, 80:413–432, 1995.

[29] D. S. Broomhead, J. P. Huke, and M. A. S. Potts. Cancelling deterministic noise by constructing nonlinear inverses to linear filters. *Physica D*, 89:439–458, 1996.

[30] P. L. Butzer. The Shannon sampling theorem and some of its generalisations: An overview. In *Constructive Function Theory (Proc. Conf. Varna, Bulgaria, 1981)*, pages 258–274. Publishing House of Bulgarian Academy of Sciences, 1983.

[31] P. L. Butzer. A survey of the Whittaker-Shannon sampling theorem and some of its extensions. *J. Math. Res. Exposition*, 3:185–212, 1983.

[32] P. L. Butzer and W. Engels. On the implementation of the Shannon sampling series for bandlimited signals. *IEEE Trans. Info. Theory*, 29:314–318, 1983.

[33] P. L. Butzer, W. Engels, S. Ries, and R. L. Stens. The Shannon sampling series and the reconstruction of signals in terms of linear, quadratic and cubic splines. *SIAM J. Appl. Math.*, 46:299–323, 1986.

[34] P. L. Butzer, S. Ries, and R. L. Stens. The Whittaker-Shannon sampling theorem, related theorems and extensions: a survey. In *Proc. JIEEEC'83*, pages 50–56, Amman, Jordan, 1983.

[35] P. L. Butzer, S. Ries, and R. L. Stens. Shannon's sampling theorem, Cauchy's integral formula, and related results. In *Anniversary volume on approximation theory and functional analysis (Proc. Conf. Math. Res. Inst. Oberwolfach, Germany)*, pages 363–377. Birkhäuser Verlag, 1984.

[36] P. L. Butzer, W. Splettstösser, and R. L. Stens. The sampling theorem and linear prediction in signal analysis. *Jahresbericht Deutscher Math.-Verein.*, 90:1–70, 1988.

[37] P. L. Butzer and R. L. Stens. Prediction of non-bandlimited signals from past samples in terms of splines of low degree. *Math. Nachr.*, 132:115–130, 1987.

[38] P. L. Butzer and R. L. Stens. Sampling theory for not necessarily bandlimited functions: an historical overview. *SIAM Review*, 34:40–53, 1992.

[39] P. L. Butzer and R. L. Stens. Linear prediction by samples from the past. In R. J. Marks II, editor, *Advanced topics in Shannon sampling and interpolation theory*, chapter 5. Springer-Verlag, 1993.

[40] P. L. Butzer, R. L. Stens, and M. Wehrens. The continuous Legendre transform, its inverse transform, and applications. *Int. J. Math. Sci.*, 3:47–67, 1980.

[41] G. Calvagno and D. Munson. New results on Yen's approach to interpolation from nonuniformly spaced samples. *Proc. ICASSP*, pages 1533–1538, 1990.

[42] L. L. Campbell. A comparison of the sampling theorems of Kramer and Whittaker. *J. SIAM*, 12:117–130, 1964.

[43] M. Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35:335–356, 1989.

[44] J. W. S. Cassels. *Local Fields.* CUP, 1986.

[45] J. W. S. Cassels. *Lectures on Elliptic Curves.* CUP, 1991.

[46] A.-L. Cauchy. Mémoire sur diverses formules d'analyse. *Comptes Rendus Paris*, 12:283–298, 1841. In French. Translation in [209].

[47] J. Cezanne and A. Papoulis. The use of modulated splines for the reconstruction of bandlimited signals. *IEEE Trans. Acoustics, Speech, Sig. Proc.*, 36:1521–1525, 1988.

[48] D. S. Chen and J. P. Allebach. Analysis of error in reconstruction of two-dimensional signals from irregularly spaced samples. *IEEE Trans. Acoustics, Speech, Sig. Proc.*, 35:173–180, 1987.

[49] K. F. Cheung. Multidimensional analogues of the sampling theorem. In R. J. Marks II, editor, *Advanced topics in Shannon sampling and interpolation theory*, chapter 3. Springer-Verlag, 1993.

[50] L. O. Chua and M. Hasler, editors. *IEEE Trans. Circuits & Systems Special Issue on Chaos in Nonlinear Electronic Circuits*, volume 40(10,11):637–884, 1993.

[51] M. E. Davies. An iterated function approximation in shadowing time series. *Phys. Lett. A*, 169:251–258, 1992.

[52] M. E. Davies. Noise reduction by gradient descent. *Int. J. Bifurcation and Chaos*, 3:113–118, 1992.

[53] C. de Boor. *A practical guide to splines.* Springer-Verlag, New York, 1978.

[54] J. de Sousa Pinto. A generalised hankel convolution. *SIAM J. Math. Anal.*, 16:1335–1346, 1985.

[55] M. M. Dodson and A. M. Silva. An algorithm for optimal regular sampling. *Sig. Proc.*, 17:169–174, 1989.

[56] F. U. Dowla and J. H. McClellan. MEM spectral analysis for nonuniformly sampled signals. *Proc. ICASSP*, pages 79–85, 1981.

[57] R. J. Duffin and A. C. Schaeffer. Some properties of entire functions of exponential type. *Bull. Amer. Math. Soc.*, 44:236–240, 1938.

[58] D. J. Christini et al. Application of linear and nonlinear time series analysis modeling to heart rate dynamics analysis. *IEEE Trans. Biomedical Eng.*, 42:411–415, 1995.

[59] P. Laguna et al. Power spectral density of unevenly sampled heart rate data. *Proc. IEEE-EMBC'95*, pages 157–158, 1995.

[60] J. D. Farmer and J. J. Sidorowich. Predicting chaotic time series. *Phys. Rev. Lett.*, 59:845–848, 1987.

[61] J. D. Farmer and J. J. Sidorowich. Optimal shadowing and noise reduction. *Physica D*, 47:373–392, 1991.

[62] H. G. Feichtinger. Iterative methods for scattered data approximation of smooth signals. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 16–31, 1995.

[63] H. G. Feichtinger and K. Gröchenig. Irregular sampling theorems and series expansions of bandlimited functions. *J. Math. Anal. Appls.*, 167:530–556, 1992.

[64] H. G. Feichtinger, K. Gröchenig, and T. Strohmer. Efficient numerical methods in non-uniform sampling theory. *Numerische Matematik*, 69:423–440, 1995.

[65] P. J. S. G. Ferreira. Incomplete sampling series and the recovery of missing samples from oversampled bandlimited signals. *IEEE Trans. Sig. Proc.*, 40:225–227, 1992.

[66] P. J. S. G. Ferreira. The stability of a procedure for the recovery of lost samples in band-limited signals. *Sig. Proc.*, 40:195–205, 1994.

[67] P. J. S. G. Ferreira. On two recent approaches to the reconstruction of signals from nonuniform samples. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 263–267, 1995.

[68] P. J. S. G. Ferreira and J. M. N. Vieira. Detection and correction of missing samples. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 169–174, 1997.

[69] G. Foster. The cleanest Fourier spectrum. *Astron. J.*, 109:1889–1901, 1995.

[70] W. A. Frank. Sampling requirements for Volterra system identification. *IEEE Sig. Proc. Lett.*, 3:266–267, 1996.

[71] A. G. García and M. A. Hernández-Medina. Finite sampling associated with second-order difference problems. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 385–390, 1997.

[72] A. G. García and M. A. Hernández-Medina. Sampling theorems associated with difference Sturm-Liouville problems. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 333–338, 1997.

[73] P. A. Glendinning. *Stability, instability and chaos: An introduction to the theory of nonlinear differential equations.* CUP, 1994.

[74] S. J. Godsill. *The restoration of degraded audio signals.* PhD thesis, Dept. of Engineering, University of Cambridge, 1993.

[75] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series and Products.* Academic, fifth edition, 1994.

[76] M. Greitans. Reconstruction of nonuniformly sampled bandpass random signal waveforms. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 290–293, 1995.

[77] R. H. Haddad and T. W. Parsons. *Digital signal processing: theory, applications and hardware.* Computer Science Press (W. H. Freeman & Co.), New York, 1991.

[78] J. M. Hammersley and D. C. Handscomb. *Monte Carlo methods.* Methuen, 1964.

[79] G. H. Hardy and E. M. Wright. *An introduction to the theory of numbers.* OUP, 1979.

[80] A. C. Harvey. *Filtering, structural time series analysis and the Kalman filter.* CUP, 1989.

[81] H. D. Helms and J. B. Thomas. Truncation error of sampling theorem expansion. *Proc. IRE*, 50:179–184, 1962.

[82] J. R. Higgins. Five short stories about the cardinal series. *Bull. Amer. Math. Soc.*, 12:45–89, 1985.

[83] J. R. Higgins. Sampling theorems and the contour integral method. *Appl. Anal.*, 41:155–169, 1991.

[84] J. R. Higgins. *Sampling in Fourier and signal analysis: Foundations.* OUP, 1996.

[85] J. R. Higgins. Sampling theorems from the iteration of low-order differential operators. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, page 463, 1997.

[86] G. Hinsen. *Abtastsätze mit unregelmäßigen Stützstellen: Rekonstruktionsformeln, Konvergenzaussagen und Fehlerbetrachtungen.* PhD thesis, RWTH Aachen & University of Bonn, 1991. In German.

[87] S. Holm, B. Elgetun, and G. Dahl. Weight- and layout- optimized sparse arrays. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 97–102, 1997.

[88] R. F. Hoskins and J. de Sousa Pinto. Sampling expansions and generalised translation invariance. *J. Franklin Inst.*, 317:323–332, 1984.

[89] P. J. Huber. *Robust Statistics.* Wiley, New York, 1981.

[90] J. P. Huke. Embedding nonlinear dynamical systems: A guide to Takens' theorem. Technical report, DRA Malvern, UK, 1993.

[91] R. J. Hyndman. Approximations and boundary conditions for threshold autoregressive processes. *J. Appl. Prob.*, 31:1103–1109, 1994.

[92] A. J. E. Janssen, R. N. Veldhuis, and L. B. Vries. Adaptive interpolation of discrete time signals that can be modeled as autoregressive processes. *IEEE Trans. Acoustics, Speech, Sig. Proc.*, 34:317–330, 1986.

[93] A. J. Jerri. Sampling expansion for the L_ν^α-Laguerre integral transform. *J. Res. National Bureau of Standards B*, 80:415–418, 1976.

[94] A. J. Jerri. The Shannon sampling theorem — its various extensions and applications: A tutorial review. *Proc. IEEE*, 65:1565–1596, 1977.

[95] A. J. Jerri. Towards a discrete Hankel transform and its applications. *Appl. Anal.*, 80:97–109, 1978.

[96] A. J. Jerri. An extended Poisson type summation formula for general integral transforms and aliasing error bound for the generalized sampling theorem. *J. Appl. Anal.*, 26:199–221, 1988.

[97] A. J. Jerri. *Integral and discrete transforms with applications and error analysis.* Marcel Dekker, New York, 1992.

[98] A. J. Jerri. Error analysis in application of generalizations of the sampling theorem. In R. J. Marks II, editor, *Advanced topics in Shannon sampling and interpolation theory*, chapter 7. Springer-Verlag, 1993.

[99] A. J. Jerri and E. J. Davis. Application of the sampling theorem to boundary value problems. *J. Eng. Math*, 8:1–8, 1974.

[100] A. J. Jerri and I. A. Joslin. Truncation error for the generalized Bessel type sampling series. *J. Franklin Inst.*, 314:323–328, 1982.

[101] R. H. Jones. Fitting a continuous time autoregression to discrete data. In D. F. Findley, editor, *Applied Time Series Analysis II.* Academic, 1981.

[102] R. H. Jones. Fitting multivariate models to unequally-spaced data. In E. Parzen, editor, *Time series analysis of irregularly observed data*, pages 158–188. Springer, 1984.

[103] M. I. Kadec. The exact value of the Paley-Wiener constant. *Soviet. Math. Dokl.*, 20:559–561, 1964.

[104] T. Kailath. *Lectures on Wiener and Kalman filtering.* Springer-Verlag, New York, 1981.

[105] V. E. Katsnelson. The mean-periodic continuation method in sampling theory for functions with multi-band spectrum. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 457–462, 1997.

[106] R. A. Katz, editor. *Chaotic, fractal and nonlinear signal processing.* AIP Press, 1995.

[107] S. M. Kay and S. L. Marple Jr. Spectrum analysis—A modern perspective. *Proc. IEEE*, 69:1380–1419, 1981.

[108] R. King, M. Ahmadi, R. Gorgui-Naguib, A. Kwabwe, and M. Azimi-Sadjadi. *Digital filtering in one and two dimensions.* Plenum, 1989.

[109] A. M. Kinghorn, P. E. Holbourn, M. J. Smith, and J. F. Roulston. The Blue Vixen Radar for Sea Harrier FRS.2. *GEC Review*, 9:131–145, 1994.

[110] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations.* Springer, 1992.

[111] I. Kluvánek. Sampling theorem in abstract harmonic analysis. *Mat. Casopis Sloven. Akad. Vied.*, 15:43–48, 1965.

[112] M. T. M. Koper. Bifurcations of mixed-mode oscillations in a three-variable autonomous van der Pol-Duffing model with a cross-shaped phase diagram. *Physica D*, 80:72–94, 1995.

[113] H. P. Kramer. A generalised sampling theorem. *J. Math. Phys.*, 38:68–72, 1959.

[114] H. J. Landau. On the recovery of a bandlimited signal after instananeous compnading and subsequent band limiting. *Bell Systems Tech. J.*, 39:351–364, 1960.

[115] H. J. Landau. Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Math.*, 117:37–52, 1967.

[116] H. J. Landau. Sampling, data transmission, and the Nyquist rate. *Proc. IEEE*, 55:1701–1706, 1967.

[117] H. J. Landau and W. L. Miranker. The recovery of distorted bandlimited signals. *J. Math. Anal. and Appl.*, 2:97–104, 1961.

[118] H. J. Landau and H. O. Pollak. Prolate spheroidal wave functions, Fourier analysis, and uncertainty II. *Bell Syst. Tech. J.*, 40:65–84, 1961.

[119] H. J. Landau and H. O. Pollak. Prolate spheroidal wave functions, Fourier analysis, and uncertainty III. *Bell Syst. Tech. J.*, 41:1295–1336, 1962.

[120] J. A. Legg. *Synthetic aperture radar using non-uniform sampling.* PhD thesis, University of Adelaide, 1997.

[121] M. J. Lighthill. *An introduction to Fourier analysis ans generalised functions.* CUP, 1958.

[122] N. R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, pages 447–462, 1976.

[123] Yu. Lyubarskii and G. Chistyakov. Reconstruction of bandlimited signals via random samples and related problems on random exponential sytems. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 44–48, 1995.

[124] Yu. Lyubarskii and W. R. Madych. The characterisation and recovery of a class of band limited signals from irregular samples. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 317–320, 1997.

[125] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63:561–580, 1975.

[126] R. J. Marks II. *Introduction to Shannon sampling and interpolation theory.* Springer-Verlag, New York, 1991.

[127] R. J. Marks II. *Advanced topics in Shannon sampling and interpolation theory.* Springer-Verlag, New York, 1993.

[128] S. L. Marple Jr. *Digital spectral analysis with applications.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[129] D. W. Marquardt and S. K. Acuff. Direct quadratic spectrum estimation with irregularly spaced data. In E. Parzen, editor, *Time series analysis of irregularly observed data*, pages 211–223. Springer, 1984.

[130] R. J. Martin. Dynamics of a double van der Pol oscillator. Seminar at Bristol University, June 1997. (The dynamics of a unijunction transistor circuit with two capacitors were similar to those described in [112].).

[131] R. J. Martin. Full-field optical Doppler anemometry. Master's thesis, Cranfield Institute of Technology, 1993.

[132] R. J. Martin. All-pole modelling of irregularly-sampled data. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 247–252, 1995.

[133] R. J. Martin. Autoregression and irregular sampling: Filtering. *Sig. Proc.*, 1996. To appear.

[134] R. J. Martin. Autoregressive modelling of irregularly-sampled data. In G. Ramponi, G. L. Sicuranza, S. Carrato, and S. Marsi, editors, *Signal Processing VIII (Proc. EUSIPCO-96)*, pages 2033–2036, 1996.

[135] R. J. Martin. Model-based approaches to signal analysis. *GEC J. Research*, 13:28–41, 1996.

[136] R. J. Martin. Autoregression and irregular sampling: Spectral estimation. *Sig. Proc.*, 1997. To appear.

[137] R. J. Martin. Functional fitting and its connections with sampling theory. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 407–412, 1997.

[138] R. J. Martin. An introduction to Shannon sampling and interpolation theory, with generalisations to nonuniform sampling. *GEC J. Tech.*, 14:19–26, 1997.

[139] R. J. Martin. Generalised prediction error approach to the spectral analysis and filtering of irregularly sampled data. In J. G. McWhirter, editor, *Proc. IMA Conf. Math. Sig. Proc. 1996.* OUP, 1998. To appear.

[140] R. J. Martin. One-sided sampling series gives fast reconstruction algorithm. *IEEE Trans. Sig. Proc.*, 1998. To appear.

[141] R. J. Martin and D. A. Castelow. Reconstruction of multiband signals using irregular sampling. *GEC J. Tech.*, 14:180–185, 1997.

[142] R. J. Martin and M. J. Kearney. Remote sea current sensing using HF radar: An autoregressive approach. *IEEE J. Oceanic Eng.*, 22:151–155, 1997.

[143] F. Marvasti. *A unified approach to zero-crossings and nonuniform sampling.* Nonuniform Publications, Oak Park, Illinois, 1987.

[144] E. Masry. Random sampling and reconstruction of spectra. *Information and Control*, 19:275–288, 1971.

[145] E. Masry. Alias-free sampling: an alternative conceptualization and its applications. *IEEE Trans. Info. Theory*, 24:317–324, 1978.

[146] E. Masry. Poisson sampling and spectral estimation of continuous time processes. *IEEE Trans. Info. Theory*, 24:173–183, 1978.

[147] E. Masry. Non-parametric covariance estimation from irregularly-spaced data. *Adv. Appl. Prob.*, 15:113–132, 1983.

[148] E. Masry and M. C. C. Lui. Discrete-time spectral estimation of continuous-parameter processes—A new consistent estimate. *IEEE Trans. Info. Theory*, 22:298–312, 1976.

[149] R. N. McDonough. Maximum-entropy spatial processing of array data. *Geophysics*, 39:843–851, 1974.

[150] R. M. Mersereau and M. J. T. Smith. *Digital filtering: A computer laboratory textbook.* Wiley, 1994.

[151] T. K. Moon. The expectation-maximization algorithm. *IEEE Sig. Proc. Mag.*, pages 47–60, November 1996.

[152] G. Nasri-Roudsari. Sampling expansions of Chebychev and Legendre functions. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 146–149, 1995.

[153] W. I. Newman. Extension to the maximum entropy method. *IEEE Trans. Info. Theory*, 23:89–93, 1977.

[154] W. I. Newman. Extension to the maximum entropy method II. *IEEE Trans. Info. Theory*, 25:705–708, 1979.

[155] A. H. Nuttall. Spectral analysis of a univariate process with bad data points v̈ia maximum entropy and linear predictive techniques. Technical Report TR-5303, Naval Underwater Systems Center, New London, Conn., 1974.

[156] H. Nyquist. Certain topics in telegraph transmission theory. *AIEE Trans.*, 47:617–644, 1928.

[157] B. Øksendal. *Stochastic Differential Equations.* Springer, 1992.

[158] E. Ott, T. Sauer, and J. A. Yorke. *Coping with Chaos.* Wiley, 1994.

[159] R. E. A. C. Paley and N. Wiener. *Fourier Transforms in the Complex Domain*, volume 19. Amer. Math. Soc. Colloq. Pub., New York, 1934.

[160] A. Papoulis. *Probability, Random variables, and Stochastic Processes.* McGraw-Hill, New York, third edition, 1991.

[161] E. Parzen. Mathematical considerations in the estimation of spectra. *Technometrics*, 3:167–190, 1961.

[162] H. O. Pollak and D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty I. *Bell Syst. Tech. J.*, 40:43–64, 1961.

[163] M. J. D. Powell. The theory of radial basis function approximation in 1990. Technical report, DAMTP, University of Cambridge, 1990.

[164] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C.* CUP, first edition, 1988. (Anything in here is also in the second edition.).

[165] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C.* CUP, second edition, 1992.

[166] M. B. Priestley. *Non-linear and non-stationary time series analysis.* Academic, 1991.

[167] J. G. Proakis. *Digital Communications.* McGraw-Hill, 1995.

[168] M. Rauth. Application of 2D methods for scattered data approximation to geophysical data sets. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 38–43, 1995.

[169] M. Rauth and T. Strohmer. A frequency domain approach to the recovery of geophysical potentials. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 109–114, 1997.

[170] M. D. Rawn. Generalized sampling theorems for Bessel-type transforms of bandlimited functions and distributions. *SIAM J. Appl. Math.*, 49:638–649, 1989.

[171] A. Requicha. The zeros of entire functions: Theory and engineering applications. *Proc. IEEE*, 68:308–328, 1980.

[172] S. Ries. Reconstruction of real and analytic band-pass signals from a finite number of samples. *Sig. Proc.*, 33:237–257, 1993.

[173] I. W. Sandberg. On the properties of some systems that distort signals—I. *Bell Syst. Tech. J.*, 42:2033–2046, 1963.

[174] I. W. Sandberg. On the properties of some systems that distort signals—II. *Bell Syst. Tech. J.*, 43:91–111, 1964.

[175] T. Sauer. Interspike interval embedding of chaotic signals. Unpublished.

[176] T. Sauer. Reconstruction of dynamical systems from interspike intervals. *Phys. Rev. Lett.*, 72:3811–3814, 1994.

[177] T. Sauer. Reconstruction of integrate-and-fire dynamics. *Fields Institute Comms.*, To appear.

[178] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65:579–616, 1991.

[179] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[180] J. D. Scargle. Studies in astronomical time series II. Statistical aspects of spectral analysis of unevenly sampled data. *Astrophys. J.*, 263:835–853, 1982.

[181] W. Schikhof. *Ultrametric Calculus.* CUP, 1984.

[182] G. Schmeisser. Sampling, Gaussian quadrature, and Poisson summation formula. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 327–332, 1997.

[183] M. A. Schonewille and A. J. W. Duijndam. Mixed Fourier-Radon reconstruction of irregularly and sparsely sampled seismic data. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 115–120, 1997.

[184] H. Schuster. *Deterministic chaos: An introduction.* VCH, Weinheim, 1989.

[185] S. C. Scoular and W. J. Fitzgerald. Periodic nonuniform sampling of multiband signals. *Sig. Proc.*, 28:195–200, 1992.

[186] K. Seip. On the connection between exponential bases and certain related sequences in $L^2(-\pi, \pi)$. *J. Functional Analysis*, 130:131–60, 1995.

[187] K. Seip. A simple construction of exponential bases in L^2 of the union of several intervals. *Proc. Edinb. Math. Soc.*, 38:171–77, 1995.

[188] C. E. Shannon. Communications in the presence of noise. *Proc. IRE*, 37:10–21, 1949.

[189] H. S. Shapiro and R. A. Silverman. Alias-free sampling of random noise. *SIAM J. Appl. Math.*, 8:225–236, 1960.

[190] T. N. Shorey and R. Tijdeman. *Exponential Diophantine equations.* CUP, 1986.

[191] R. H. Shumway. Some applications of the EM algorithm to analysing incomplete time series data. In E. Parzen, editor, *Time series analysis of irregularly observed data*, pages 290–324. Springer, 1983.

[192] D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty IV. *Bell Syst. Tech. J.*, 43:3009–3057, 1964.

[193] D. Slepian. On bandwidth. *Proc. IEEE*, 64:292–300, 1976.

[194] D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty V. *Bell Syst. Tech. J.*, 57:1371–1430, 1978.

[195] L. A. Smith. Identification and prediction of low-dimensional dynamics. *Physica D*, 58:50–76, 1992.

[196] W. Splettstösser. 75 years' aliasing error in the sampling theorem. *Proc. EUSIPCO*, pages 1–4, 1983.

[197] H. Stark. Sampling theorems in polar and spiral coordinates. In R. J. Marks II, editor, *Advanced topics in Shannon sampling and interpolation theory*, chapter 6. Springer-Verlag, 1993.

[198] J. Stark. Delay embeddings for forced systems: I. Deterministic forcing. Submitted.

[199] J. Stark. Recursive prediction of chaotic time series. *J. Nonlinear Science*, 3:197–223, 1993.

[200] J. Stark. Adaptive model selection using orthogonal least-squares methods. *Proc. Roy. Soc. London A*, 453:21–42, 1997.

[201] J. Stark. Invariant graphs for forced systems. *Physica D*, 109:163–179, 1997.

[202] J. Stark, D. S. Broomhead, M. E. Davies, and J. P. Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis*, 8:5303–5314, 1997.

[203] J. Stark and M. E. Davies. Recursive filters driven by forced systems. *IEE Digest*, 143(5):1–16, 1997. IEE Colloquium on Exploiting Chaos in Signal Processing.

[204] I. Stewart. *Galois Theory.* Chapman and Hall, 1982.

[205] I. Stewart and D. Tall. *Complex Analysis.* CUP, 1983.

[206] D. C. Stickler. An upper bound on aliasing error. *Proc. IEEE*, 55:418–419, 1967.

[207] T. Strohmer. *Efficient methods for digital signal and image reconstruction from nonuniform samples.* PhD thesis, University of Vienna, 1993.

[208] T. Strohmer. Computationally sttractive reconstruction of bandlimited images from irregular samples. *IEEE Trans. Image Proc.*, 6:540–548, 1997.

[209] B. Summers. *On extensions to the sampling theorem and processing techniques for non-equispaced sampled-data systems.* PhD thesis, Polytechnic of Central London [now University of Westminster], 1988.

[210] F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L. S. Young, editors, *Dynamical systems and turbulence (LMS 898).* Springer, 1980.

[211] A. Tarczynski and G. D. Cain. Reliability of signal reconstruction from finite sets of samples. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, pages 181–186, 1997.

[212] J. M. T. Thompson and H. B. Stewart. *Nonlinear dynamics and chaos.* Wiley, 1986.

[213] H. Tong and I. Yeung. Threshold autoregressive modelling in continuous time. *Statistica Sinica*, 1:411–430, 1991.

[214] J. Tsimbinos. *Identification and compensation of nonlinear distortion.* PhD thesis, School of Electronic Engineering, University of South Australia, 1995.

[215] J. Tsimbinos and K. V. Lever. Sampling frequency requirements for identification and compensation of nonlinear systems. *Proc. ICASSP*, pages 513–516, 1994.

[216] J. Tsimbinos and K. V. Lever. The computational complexity of nonlinear compensators based on the Volterra inverse. In *Statistical Signal and Array Processing (Proc. IEEE SSAP-96, Corfu, Greece, June 1996)*, pages 387–390, 1996.

[217] T. J. Ulrych and R. W. Clayton. Time series modelling and maximum entropy. *Phys. Earth Planet. Inter.*, 12:188–200, 1976.

[218] B. van der Pol. The heartbeat considered as a relaxation oscillator, and an electrical model of the heart. *Phil. Mag.*, 7:763–775, 1928.

[219] J. J. Voss. A contribution to nonuniform sampling with complex nodes. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 247–252, 1995.

[220] J. J. Voss. A sampling theorem for bandlimited L^p functions. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'97, Aveiro, Portugal, June 1997)*, 1997.

[221] G. Watson. *Treatise on the theory of Bessel functions.* CUP, second edition, 1962.

[222] P. Weiss. Sampling theorems associated with Sturm-Liouville systems. *Bull. Amer. Math. Soc.*, 63:242, 1957.

[223] D. J. Wingham. The reconstruction of a band-limited function and its Fourier transform from a finite number of samples at arbitrary locations by singular value decomposition. *IEEE Trans. Sig. Proc.*, 40:559–570, 1992.

[224] K. Yao and J. B. Thomas. On truncation error for sampling representations of bandlimited signals. *IEEE Trans. Aero. Electron. Syst.*, 2:640–647, 1966.

[225] K. Yao and J. B. Thomas. On some stability and interpolatory properties of nonuniform sampling expanions. *IEEE Trans. Circuit Theory*, 14:404–408, 1967.

[226] J. L. Yen. On nonuniform sampling of bandwidth-limited signals. *IRE Trans. Circuit Theory*, 3:251–257, 1956.

[227] G. U. Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Phil. Trans. Roy. Soc. London A*, 226:267–298, 1927.

[228] A. I. Zayed. Sampling expansion for the continuous Bessel transform. *Appl. Anal.*, 27:47–65, 1988.

[229] A. I. Zayed. *Advances in Shannon's sampling theory.* CRC Press, 1993.

[230] Y. M. Zhu. Generalised sampling theorem. *IEEE Trans. Circuits Systems II*, 39:587–588, 1992.

[231] M. Zwaan. Bounds for errors in nonuniform sinc interpolation. In *Sampling Theory and Applications (Proc. Workshop SAMPTA'95, Jurmala (Riga), Latvia, September 1995)*, pages 55–58, 1995.