Lecture 14

• Spectral analysis revisited
  – Comments on frequency localization
  – Lomb-Scargle method for data with gaps
    • preserves exponential statistics and statistical independence
  – Red noise, leakage, missing information
  – Entropy and information

Reading: Chapter 13 of Gregory “The Bayesian Revolution in Spectral Analysis”

DFT Examples

• Revisit spectral windows briefly
Comments on Windows

• The rectangular windows in the gap case are wider than for the randomly sampled case
  – i.e. view the “data-on” segments as corresponding to a rectangle function
• The uncertainty principle says that a wider rectangle in one domain corresponds to a narrower sinc function in the conjugate domain
• But the spectral windows show the opposite
• Why?

Comments:

1. Gaps increase the noise floor and lowers the signal level so lowers the S/N.
2. Gaps broaden the spectral lines.
3. Gaps distort the shapes of spectral lines.

Sampling windows and analogies:

We can think of the uniformly sample time series as an infinite time series multiplied by a rectangular window of duration $T$.

The cases with gaps and random sampling represents further windowing.

Generally we can write the data as the product of the infinite time series and a window function.

$$y(t) = y_\infty(t) \times W(t)$$

Then the power spectrum is

$$S_y(f) = S_{y_\infty}(f) \ast |\tilde{W}(f)|^2.$$ 

So the spectral window $|\tilde{W}(f)|^2$ is the resolution function or ‘point spread function’ for the spectrum.

By analogy with interferometric imaging, this can also be called the ‘dirty beam.’

The DFT-based spectrum by analogy is the ‘dirty spectrum.’

The CLEAN algorithm is a restoration algorithm that strives to deconvolve the dirty beam from the dirty spectrum to produce a CLEANed spectrum.
Figures 5 and 6 show the window functions for the gapped time series and the randomly sampled time series, respectively.

The window function for the gapped time series has wider intervals of zeros than does the case with random sampling. Each of these can be thought of as a sequence of rectangle functions with variable widths and locations along the time axis.

*Question:* The average rectangle duration of the randomly-sample case is much shorter than the typical gap width in the case with gap blocks. Yet, the width of the respective window functions in the frequency domain is smaller for the randomly-sampled case. Isn’t this counter to the uncertainty principle?

No!

For each case, let

- \( N_r \) = number of rectangles in a window function.
- \( t_j \) = centroid time of each rectangle.
- \( w_j \) = width of each rectangle.

Let’s construct the Fourier transform of the window function.

Start with

\[
\Pi(t) \iff \text{sinc } f
\]

By the scaling theorem we have

\[
\frac{1}{w_j} \Pi \left( \frac{t - t_j}{w_j} \right) \iff \text{sinc } fw_j
\]

Using the shift theorem

\[
\frac{1}{w_j} \Pi \left( \frac{t - t_j}{w_j} \right) \iff e^{-2\pi i f_j} \text{sinc } fw_j
\]

Summing over all rectangles, we have

\[
W(t) = \sum_{j=1}^{N_r} \frac{1}{w_j} \Pi \left( \frac{t - t_j}{w_j} \right) \iff \text{sinc } fw_j \sum_{j=1}^{N_r} e^{-2\pi i f_j}
\]

The \( \text{sinc } fw_j \) function is indeed wider in frequency if \( w_j \) is narrower.

However the sum of phase factors will average to zero in the limit \( N_r \to \infty \) and with a deviation from zero \( \sim 1/\sqrt{N_r} \) except for \( f = 0 \) when the sum equals \( N_r \).

This is why the frequency domain window function is narrower for the randomly sampled case than for the gapped case.
A Bayesian Approach to Spectral Analysis

Chirped Signals
Chirped signals are oscillating signals with time variable frequencies, usually with a linear variation of frequency with time. E.g.,

\[ f(t) = A \cos(\omega t + \alpha t^2 + \theta). \]

Examples:
- Plasma wave diagnostic signals
- Signals propagated through dispersive media (seismic cases, plasmas)
- Gravitational waves from inspiraling binary stars
- Doppler-shifted signals over fractions of an orbit (e.g., acceleration of pulsar in its orbit)

Jaynes’ Approach to Spectral Analysis:
Briefly by Gregory in Chapter 13 of Bayesian Logical Data Analysis for the Physical Sciences
Result: Optimal processing is a nonlinear operation on the data without recourse to smoothing. However, the DFT-based spectrum (the “periodogram”) plays a key role in the estimation.

Fresnel Function and its DFT

Fresnel function
\[ c(t) = e^{i\alpha t^2} \]

What is the FT of \( c(t) \)?
Define
\[ C(\omega, \alpha) \equiv N^{-1}(P^2 + Q^2) = N^{-1} \sum_{t} \sum_{t'} y(t)y(t') \cos[\omega(t - t') + \alpha(t - t')] \, . \]

Then the integral over \( \theta \) gives
\[
\int_{0}^{2\pi} d\theta L(A, \omega, \alpha, \theta) \equiv I_0 \left( \frac{A_1 NC(\omega, \alpha)}{\sigma^2} \right)
\]
and the marginalized likelihood is
\[
L(A, \omega, \alpha) = e^{-\frac{NA^2}{4\sigma^2}} I_0 \left( \frac{A_1 NC(\omega, \alpha)}{\sigma^2} \right).
\]

**Interpretation of the Bayesian and Fourier Approaches**

We found the marginalized likelihood for the frequency and chirp rate to be
\[
L(A, \omega, \alpha) = e^{-\frac{NA^2}{4\sigma^2}} I_0 \left( \frac{A_1 NC(\omega, \alpha)}{\sigma^2} \right)
\]
and the limiting form for the Bessel function’s argument \( x \gg 1 \) is
\[
I_0(x) \sim \frac{e^x}{\sqrt{2\pi x}}
\]
In this case the marginalized likelihood is
\[
L(A, \omega, \alpha) \propto e^{-\frac{NA^2}{4\sigma^2}} I_0 \left( \frac{A_1 NC(\omega, \alpha)}{\sigma^2} \right)
\]
\[
\propto e^{-\frac{NA^2}{4\sigma^2}} \times e^{-\sigma^2 \left( \frac{(2\pi A_1 NC(\omega, \alpha))}{\sigma^2} \right)^2}
\]
Since \( C(\omega, \alpha) \) is large when \( \omega \) and \( \alpha \) match those of any true signal, we see that it is exponen-
tiated as compared to appearing linearly in the periodogram.
Comparison of Spectral Line Localization Properties

Claim: While the periodogram gives a spectral line that is much broader than the width of the posterior PDF for frequency, the ability to localize the spectral line in frequency is the same for both approaches.

Periodogram: The signal-to-noise ratio (S/N) of the line is \( \sim NA/\sigma \) (as in DFT of complex exponential). The spectral resolution is \( \delta \omega_{\text{res}} \sim 2\pi/(2T+1) \) (since our time interval is \([-T, T]\)). The width of the line (e.g. FWHM) is of order the spectral resolution.

Assume S/N is large.

Posterior PDF: The PDF for \( \omega \) is dominated by the exponential factor

\[
E(\omega) = \exp\{A_N C(\omega, 0)/\sigma^2\}
\]

From the expression for \( C \) we have \( C_{\text{max}} = C(\omega = \omega_0) = NA^2/4 \) so

\[
E_{\text{max}} = e^{A_N C_{\text{max}}/\sigma^2} = e^{NA/\sigma^2}
\]

For offset frequencies \( \omega = \omega_0 + \delta \omega \) we can expand various things to show that

\[
E(\omega_0 + \delta \omega) \approx E_{\text{max}} e^{(NA/\sigma^2)(\delta \omega + (2T+1)/N\pi)^2}
\]

This function has a width when the exponential = 1/2 or \( \delta \omega = \sqrt{12} \frac{\sigma}{NA} \)

In terms of resolution units this is

\[
\frac{\delta \omega_{\text{res}}}{\delta \omega} = \sqrt{12} \left( \frac{\sigma}{NA} \right) \frac{1}{S/N \text{ of line in periodogram}}
\]
Lomb-Scargle Method

- Two main features of the periodogram for uniformly distributed data:
  - the exponential PDF for spectral amplitudes
  - the independence of the spectral amplitudes \( S(f) \) and \( S(f') \) as \( T \to \infty \)
- The second property is related to the fact that \( e^{i\omega t} \) for the \( \omega \) of interest are a set of orthogonal basis functions
- With nonuniform sampling, orthogonality no longer applies

Scargle’s Approach

*Astrophysical Journal, 263, 835, 1982*

The periodogram is then conventionally defined as

\[
P_0(\omega) = \frac{1}{N_0} |\text{FT}_x(\omega)|^2
\]

\[
= \frac{1}{N_0} \left| \sum_{j=1}^{N} X_j \exp \left(-i\omega t_j \right) \right|^2
\]

\[
= \frac{1}{N_0} \left[ \left( \sum_j X_j \cos \omega t_j \right)^2 + \left( \sum_j X_j \sin \omega t_j \right)^2 \right]
\]

A redefined form preserves the properties of the uniformly sampled case:

\[
P_x(\omega) = \frac{1}{2} \left[ \sum_j X_j \cos \omega (t_j + \tau) \right] \left[ \sum_j \cos^2 \omega (t_j + \tau) \right] + \left[ \sum_j X_j \sin \omega (t_j + \tau) \right] \left[ \sum_j \sin^2 \omega (t_j + \tau) \right] \]

(10)

where \( \tau \) is defined by

\[
\tan (2\omega \tau) = \left( \sum_j \sin 2\omega t_j \right) / \left( \sum_j \cos 2\omega t_j \right) \]

(11)
LS Spectral Analysis

• Lomb got the same result by considering least-squares fitting of sinusoids

• The same issues of sidelobes, statistical errors still apply and have to be dealt with

• The attractiveness of the LS method is that it yields well-understood features of the spectral estimator

• There is an advantage to nonuniform sampling: aliasing is related to the shortest time interval between samples

Window functions derived from the time-series sampling for astrometric data of stars

equal vs unequal sampling
Spectral analysis of unevenly spaced climatic time series using CLEAN: signal recovery and derivation of significance levels using a Monte Carlo simulation

D. Heslop a, *, M.J. Dekkers a, b

a Paleomagnetic Laboratory 'Fort Hoofddijk', Faculty of Earth Sciences, Utrecht University, Budapestlaan 17, 3584 CD Utrecht, The Netherlands
b CEREGE, BP80, 13545 Aix en Provence cedex 4, France

Abstract

We present a Monte Carlo based method for the determination of errors associated with frequency spectra produced by the CLEAN transformation of Roberts et al. (1987). The Monte Carlo procedure utilises three different types of simulation involving a data stripping operation and the addition of white and red noise to the analysed time series. The simulations are tested on both synthetic and real data sets demonstrating the ability of the procedures to extract coherent information from time series characterised by the low signal-to-noise-ratio that is typical of many palaeoclimatic records. Significance levels derived for the Monte Carlo spectra of four time series from the Vostok ice core are utilised in the study of eccentricity components contained within the palaeoclimatic archive since \( \sim 420 \) ka. Inversion of the Vostok frequency spectra into the time domain reveals the differing influence of orbital parameters in the palaeoclimatic proxy records as well as the relative magnitudes of the eccentricity components contained in the time series of greenhouse gas concentration, ice volume and local temperature.

© 2002 Elsevier Science B.V. All rights reserved.

Keywords: CLEAN transformation; Unevenly spaced time series; Palaeoclimatic
Entropy and Information

Core tenet of Maximum Entropy methods:

Out of all possible (hypotheses | PDFs | …) that agree with constraints, choose the one that is maximally non-committal with respect to missing information.
Measures of Information

Background:

(1) Entropy in statistical mechanics is related to the probability that a system of some sort
arranges itself into some configuration.

(2) For systems composed of a large number of particles the most probable one is that
which maximizes the entropy subject to constraints on the system (such as conservation of
energy, conservation of particle number, etc.).

(3) For other systems (outcomes of experiment) entropy is again associated with the prob-
abilities of the various outcomes.

(4) Information systems may be viewed in a statistical sense wherein messages are quanti-
fied according to their probability of occurrence (rather than their semantic meaning).

c.f. The Mathematical Theory of Communication, Shannon & Weaver 1949

Information and Entropy: A quantitative measure

Associate information with the frequency of occurrence of an event, not the meaning of the
message.

Less probable events carry more information, e.g. events in an event space ↔ “messages”

The sun will ‘rise’ today. \( P = 1 \) : no information

It will rain today. \( P \sim \frac{1}{2} \) in Ithaca \( \Rightarrow \) some information

The sun will supernova. \( P \ll 1 \) (Astrophysics is wrong!)

Quantitative measures of information: Let \( I \) be a function of the probability of an event:

\[ I = f(P) \]

The information measure depends on the sample space of allowed messages.
To portray the situation, we must allow messages to be events in an event space that includes
ourselves and our knowledge prior to receiving a message. e.g.

\[ P\{\text{oil shortage} \mid t < 1973\} \ll 1 \]

\[ P\{\text{oil shortage} \mid t = 2011\} \not\ll 1. \]

The message “there will be an oil shortage” is much less surprising to us today given the events
of 1973-74, 1979, 1990, and the last decade. This decade there seems to be an oil glut. What
about the next one?
Example 1: Rolling a die
(Text Sec 8.4)

• Consider a weighted die:
  Side with “i” dots appears with probability $p_i$
• Fair die: sum of $i.p_i = 3.5$
• Constraint: in 10 rolls, mean outcome = 4.

• Evaluate the multiplicity for each hypothesis that satisfies the constraint.
• The one with the largest multiplicity is the one we would consider most probable.
• Aside: terms like $(N!)$ enter into calculation:
  Stirling’s approximation.
Desired features:
1. information a positive measure $f(P) \geq 0$ for $0 \leq P \leq 1$
2. $\lim_{P \to 0} f(P) = 0$
3. $f(P_1) > f(P_2)$ if $P_1 < P_2$ (less probable $\implies$ more information)
4. For statistically independent events the information should add
   $$I_{12} = I_1 + I_2 = f(P_1) + f(P_2)$$
   but the joint event of 1 and 2 occurring has probability $P_{12} = P_1 P_2$ (if independent). Therefore,
   $$I_{12} = f(P_{12}) = f(P_1 P_2) = f(P_1) + f(P_2)$$
To satisfy these conditions logarithmic functions are optimal
$$f(P) = -\log_b P$$

Units are determined by the base:
- $b = 2$ bits
- $e$ nats
- $10$ Hartley or decit

We will use $b = 2$ and $b = e$.

For a given event with probability $P$, we write $I(P) = -\log P$.
We can view $I(P)$ as a random variable if we view it as a mapping from event space to the real number line (event $\zeta \mapsto I(\zeta) = -\log P$).

The expected value over a complete set of discrete events (an ensemble) is:
$$\langle I \rangle = \sum_i P_i I(P_i) = -\sum_i P_i \log P_i$$
$\langle I \rangle$ is the mean information per event and is called the entropy of the set.
$H \equiv \langle I \rangle$ is the entropy measure over the whole set of events or messages.
Entropy and information are associated with uncertainty.
Suppose one is waiting to get a message. If one knows a priori that a particular message will occur with unit probability, then there is no uncertainty and $H = 0$.
However, if any one of many messages is possible, then $H \neq 0$ and one is more uncertain as to what the message will be.
Binary events:
\[ p + q = 1 \quad (p = \text{she loves me}; \quad q = \text{she loves me not}) \]

\[ H = -[p \log p + (1 - p) \log (1 - p)] \]

\[ H_{\text{max}} = \log 2 = -\log \left(\frac{1}{2}\right) \] corresponds to the largest uncertainty.

**n events:** \( H \) is maximized when all events are equiprobable, if there are no other constraints:
\[ P_j = \frac{1}{n}, \quad j = 1, n \]

With constraints, we can derive the maximum entropy probabilities as a constrained maximization problem: maximize \( H \) subject to the normalization constraint (about as simple as it gets)
\[ \sum_j P_j = 1 \]
\[ H = -\sum_j P_j \log P_j \]

Define
\[ J = H + \lambda \left( \sum_j P_j - 1 \right), \quad \lambda = \text{Lagrange multiplier} \]

\[ \frac{\partial J}{\partial P_i} = \frac{\partial H}{\partial P_i} + \lambda \sum_j \frac{\partial P_j}{\partial P_i} \]
\[ = -\sum_j \left\{ \delta_{ij} \log P_j + P_j \frac{\partial \log P_j}{\partial P_i} - \lambda \frac{\partial P_j}{\partial P_i} \right\} \]
\[ = -\sum_j \left\{ \delta_{ij} \log P_j + \delta_{ij} - \lambda \delta_{ij} \right\} \]
\[ = -\log P_i - 1 + \lambda = 0 \]

Therefore \( \log P_i = \lambda - 1 \) and since \( \lambda \) is the same for all \( i, P_i = \text{constant} \).

Formally, we plug back into the constraint equation (normalization)
\[ \sum_i P_i = 1 \implies nP_i = 1 \text{ or } P_i = \frac{1}{n} \]

and \( \log P_i = \lambda - 1 \implies \lambda = 1 + \log \frac{1}{n} \)
\[ \lambda = 1 - \log n \]
Image Reconstruction

- Synthesis imaging: CLEAN vs Maximum entropy.

Two very different approaches, but both effectively fill in missing information in the Fourier domain ⇔ de-convolution in the image plane.
- Consider Figures 8.4, 8.5 in text:
  Filling in missing information after removing 50%, 95%, 99% of pixels.
  - Compare different methods?
  - Interesting for a project, maybe?

Entropy of continuous RVs

For a 1st order (univariate) PDF we have

\[ H = - \int dx f_X(x) \log f_X(x) \]

where you can see that \( dx f_X(x) \) is a probability but the rest of the integrand is the logarithm of the PDF (probability per unit \( x \)). Thus the units of \( H \) depend on the particular variable used and is not invariant to a coordinate transformation.

For an nth order (multivariate) PDF,

\[ H_X = - \int d^x f_X(x) \log f_X(x) \]

\( H \) is a relative entropy because it is defined in terms of probability densities. Consequently, \( H \) is relative to the coordinate system.
Consider a coordinate transformation

\[ \mathbf{y} = A \mathbf{x} \]

where \( \mathbf{y}_i = \sum_{j=1}^n a_{ij} x_j \)

which has a Jacobian

\[ J = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{vmatrix} = |a_{ij}|^{-1} \]

\[ H_Y = \int d\mathbf{y} f_Y(\mathbf{y}) \log f_Y(\mathbf{y}) \]

\[ H_Y = H_X - \log |J| \]

\[ H_Y = H_X - \log |a_{ij}|^{-1} \]

\[ H_Y = H_X + \log |a_{ij}| \]

Generally \( H_Y \) depends on the Jacobian but for rotations, the Jacobian is unity, so \( |a_{ij}| = 1 \Rightarrow H_Y = H_X \)

---

**Simple 1-d case:** recall the PDF transformation

\[ f_y(y) = f_x(x(y)) \left| \frac{dx}{dy} \right| \]

if \( y = g(x) \) is a single valued function. Therefore

\[ H_Y = -\int dy f_y(y) \log f_y(y) - \int dy \frac{f_x(x(y))}{|dx/dy|} \left( \log f_x(x(y)) - \log |dx/dy| \right) \]

Now convert back to an integration over \( x \)

\[ y = g(x), \quad dy = \left| \frac{dg(x)}{dx} \right| dx \]

\[ H_Y = -\int dx f_x(x) \log f_x(x) \left( 1 + \frac{1}{|dx/1|} \right) \frac{dx}{|dx/1|} \]

\[ H_Y = \int f_x(x) \log f_x(x) + \int dx f_x(x) \log \left| \frac{dx}{dx} \right| \]

\[ H_Y = H_X + \int \log \left| \frac{dg}{dx} \right| dx \]

\[ H_Y = H_X + \log \left| \frac{dg}{dx} \right| \]
Example: suppose $Y = ax$ [a simple scale change], then $dy/dx = a$ and

$$\Rightarrow H_Y = H_X + \log a$$

Proof:

$$f_y(y) = \frac{1}{|a|} f_x\left(\frac{y}{a}\right)$$

$$H_y = -\int dy f_y(y) \log f_y(y)$$

$$= -\frac{1}{|a|} \int dy f_x\left(\frac{y}{a}\right) \log f_x\left(\frac{y}{a}\right) - \log |a|$$

Back to the integration over $x$ int: $dy = a \, dx$

$$= -\frac{a}{|a|} \int dx f_x(x) \log f_x(x) - \log |a| = -\frac{a}{|a|} H_x + \frac{a}{|a|} \log |a|$$

Random variable with a constraint on its variance:

Find the PDF by maximizing entropy

Let’s consider a more interesting problem: what is the PDF of an RV with specified variance if that PDF has maximum entropy?

Maximize $H = - \int dx f(x) \log f(x)$ subject to the constraints

i) $\langle x^2 \rangle = \int dx x^2 f(x) = \sigma^2 = \text{constant}$

ii) $\int dx f(x) = 1$

Therefore we maximize

$$J = H + \lambda \langle x^2 \rangle + \mu \int dx f(x) + \text{constant}$$

$$= \int dx f(x) [-\log f(x) + \lambda x^2 + \mu] + \text{constant}$$

with respect to variations in $f(x)$, with $\lambda$ and $\mu$ being Lagrange multipliers.
The $f(x)$ with maximum entropy is that for which $\delta J = 0$ for any infinitesimal change in $f(x)$:

$$\delta J = \int dx \delta f(x) \left[ -\log f(x) + \lambda x^2 + \mu \right] + f(x) \left[ -\frac{\delta f(x)}{f(x)} \right]$$

$$= \int dx \delta f(x) \left[ -\log f(x) + \lambda x^2 + \mu - 1 \right]$$

$$= 0$$

Class: you should show that this procedure yields a maximum by considering the second derivative of $J$ and show that $\partial^2 J/\partial f^2 < 0$. Since $\delta J = 0$ for any $\delta f(x)$, we have

$$\Rightarrow -\log f(x) + \lambda x^2 + \mu - 1 = 0$$

or

$$f(x) = e^{\mu-1}e^{\lambda x^2}.$$\[\text{Now plugging back into constraints 1 and 2 we “rediscover” the Gaussian PDF:}\]

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-x^2/2\sigma^2}.$$\[\text{Recall Boltzmann’s } H \text{ theorem:}\]

$$\frac{1}{2} x^2 = E = \text{energy} \quad \text{with } \sigma^2 = KT.$$\[\text{Assumption of zero values from an entropy point of view}\]

If we (effectively) assume that data values are identically zero outside the actual data span, then the values are assigned zero value with unit probability (since we are not allowing missing values to vary across an ensemble). Thus, if $P_{jk}$ = probability of the j-th sample taking on the k-th (discrete) value, then for that value

$$H = -\sum_k P_{jk} \ln P_{jk} = -P(\text{zero}) \ln P(\text{zero})$$

$$= -1 \ln 1$$

$$= 0$$

Thus, $H = 0$ ⇒ complete certainty about values of missing numbers.

Maximum entropy techniques let missing data be maximally uncertain while being consistent with the known data points.

This is the basis for maximum entropy spectral estimators.