Lecture 13

- Entropy and Information
- Missing data and entropy
- Maximum entropy spectral estimation

Reading:

Chapter 8: Maximum Entropy Probabilities (Gregory)
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 probabilities of the various outcomes.

(4) Information systems may be viewed in a statistical sense wherein messages are quantified 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 $\Leftrightarrow$ “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.

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

$$P\{\text{oil shortage} \mid t = 2011\} \ll 1.$$ 

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. $\lim_{P \to 1} f(P) = 0$

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,

$$I_{12} = f(P_{12}) = f(P_1P_2) = f(P_1) + f(P_2)$$

The only functions that satisfy these conditions are logarithmic ones

$$f(P) = - \log_b P$$

Units are determined by the base:

$$b = 2 \quad \text{bits}$$
$$e \quad \text{nats}$$
$$10 \quad \text{Hartley or decit}$$

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.

$H \equiv \langle I \rangle$ = 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:

\[ p + q = 1 \quad (p = \text{she loves me}; \, q = \text{she loves me not}) \]

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

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

**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)

\[ \sum_j P_j - 1 = 0 \]

\[ H = -\sum_j P_j \log P_j \]
Define

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

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

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)

\[
\sum_i P_i = 1 \Rightarrow nP_i = 1 \quad \text{or} \quad P_i = \frac{1}{n}
\]

\[
\text{and} \quad \log P_i = \lambda - 1 \Rightarrow \lambda = 1 + \log \frac{1}{n} \\
\lambda = 1 - \log n
\]
Ex 2: Blue-eyed Left-handed Kangaroos (Text Sec 8.8.1)

• Consider strange kangaroos:
  1/3 blue-eyed; 1/3 left-handed. Joint probability?
• Assign probability $p_i$ ($i=1..4$) to each case:
  ($BL$, $B'L$, $BL'$, $B'L'$)
• Perfect correlation: $p_1 = 1/3$.
• Perfect anti-correlation: $p_1 = 0$.
• No correlation: $p_1 = 1/9$.
  In the absence of other info, this is preferred!
• Different variation functions are possible:
  only ($p_i \log p_i$) produces the preferred result.

* 2-dimensional cases: Image reconstruction.
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,

$$ H_X = - \int d\vec{x} \, f_{\vec{X}}(\vec{x}) \log f_{\vec{X}}(\vec{x}) $$

$H$ 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
\[ \vec{y} = A \vec{x} \]  
\[ \text{e.g. } y_1 = \sum_{j=1}^{n} a_{1j} x_j \]

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

\[ H_Y = - \int d\vec{y} f_{\vec{y}}(\vec{y}) \log f_{\vec{y}}(\vec{y}) \]
\[ = H_X - \langle \log J \rangle \]
\[ = H_X - \langle \log |a_{ij}|^{-1} \rangle \]
\[ = H_X + \log |a_{ij}| \]

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

\[ f_Y(y) = \frac{f_X[x(y)]}{|dg/dx|}, \]

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

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

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

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

\[
H_Y = -\int dx \left| \frac{dg(x)}{dx} \right| f_X(x) \left\{ \log f_X(x) - \log \left| \frac{dg}{dx} \right| \right\}
\]

\[
= -\int dx f_X(x) \log f_X(x) + \int dx f_X(x) \log \left| \frac{dg}{dx} \right| + \langle \log \left| \frac{dg}{dx} \right| \rangle
\]

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

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

Proof:

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

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

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

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

$$- \frac{a}{|a|} \int dx f_X(x) \left[ \log f_X(x) - \log |a| \right] = -\frac{a}{|a|} H_x + \frac{a}{|a|} \log |a|$$
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

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

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

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

$$f(x) = e^{\mu - 1} e^{\lambda x^2}.$$ 

Now plugging back into constraints 1 and 2 we “rediscover” the Gaussian PDF:

$$e^{\mu - 1} = \frac{1}{\sqrt{2\pi \sigma^2}}, \quad \lambda = 1/2\sigma^2$$

$$f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-x^2/2\sigma^2}$$

Recall Boltzmann’s $H$ theorem:

$$\frac{1}{2}x^2 = E = \text{energy} \quad \text{with} \quad \sigma^2 = KT.$$
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 kth (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, \( H = 0 \) \( \Rightarrow \) 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.
Aspects of Spectral Analysis

The outcome of any attempt to estimate the power spectrum should be assessed through consideration of the following:

(1) confidence intervals
(2) resolution
(3) handling of gaps (missing data)
(4) bias (particularly for processes with steep power laws)

Numbers (2) and (3) are related because resolution is determined by the length of the data set $T$ and one can view data after the interval $[0, T]$ as missing data.
Linear/fixed window/Finite F.T. methods vs. Data Adaptive Methods

Spectral estimators based on the Fourier transform have the properties:

\[
\begin{align*}
\text{no smoothing} & : \\
& 1. \text{frequency resolution} : \delta f = \frac{T^{-1}}{} \quad T = \text{data span} \\
& 2. \text{estimation error} : \varepsilon = \frac{\text{Var} \hat{S}}{\langle \hat{S}(\omega) \rangle} = 1 \\
\text{with smoothing} & : \\
& 3. \text{tradeoff between } \varepsilon \text{ and } \delta f : \\
& \quad \text{spectral window } \tilde{W} \Rightarrow \varepsilon \text{ decreases, } \delta f \text{ increases,} \\
& \quad \text{sidelobes altered } \varepsilon = \frac{2}{N_{d.o.f.}}^{1/2}
\end{align*}
\]

4. Since \( \hat{R}_w(\tau) \Leftrightarrow \hat{S}_w(\omega) \), the estimator is \textbf{linear} in the correlation function estimates, \( \hat{R}_w \).

5. Also, once chosen by the analyst, the window function is \textbf{fixed} with respect to the data.

Therefore, even though the spectral estimation may be \textbf{iterative} whereby the analyst chooses by trial an \textbf{optimum} window, this optimization is performed according to intuitive/aesthetic/prior knowledge prejudice. \textit{The form of the window is not an integral part of the estimation process.}

Hence, the characterization of finite F.T. estimators as \textit{linear, fixed window, empirical} estimators.
**Different Views on Missing Data in a Time Series**

1. Transform (in the general sense) data using fixed window analyses; then correct the data for the effects of gaps in data, etc. (e.g. can apply a CLEAN algorithm to the spectrum).

vs.

2. Transform the data without making assumptions about missing data (maximum entropy techniques are “maximally noncommittal” about missing data).

‘Empirical methods such as Blackman-Tukey, which do not invoke even a likelihood function, are useful in the preliminary, exploratory phase of a problem where our knowledge is sufficient to permit intuitive judgments about how to organize a calculation (smoothing, decimation, windows, prewhitening, padding with zeroes, etc.) but insufficient to set up a quantitative model which would do the proper things for us automatically and optimally.’ [In abstract of ”On the Rationale of Maximum Entropy Methods”, E.T. Jaynes, 1982, IEEE Proc., 70, 939.]
I. Fourier transform + deconvolution approach:

Suppose time series has gaps:

The effective window on the data:

We can write sampled data as $x_s(t) = \Omega(t)x(t)$

Subtract mean and put zeroes into the gaps:

We want an estimate for the power spectrum of $x(t)$ but we can only use $x_s(t)$ in our estimates.

**FOURIER TRANSFORM:**

$$\tilde{X}_s(f) = \tilde{\Omega}(f) \ast \tilde{X}(f)$$

$$|\tilde{X}_s(f)|^2 = |\tilde{\Omega}(f) \ast \tilde{X}(f)|^2$$

**DECONVOLUTION OF THE WINDOW FUNCTION:**

Suppose $\tilde{X}(f)$ has a spectral line, then $|\tilde{\Omega}(f)|^2$ is convolved with the spectral line, adding large sidelobes (because of discontinuities).

These can be removed, to some extent, by deconvolving the $|\tilde{\Omega}(f)|^2$ from the spectral estimate.

For “simple” spectral estimates composed of a few spectral lines, the deconvolution process becomes, essentially, a subtraction.
For example a spectral estimate may appear as

and the window function has $|F.T.|^2$ of

By subtracting $\gamma|\tilde{\Omega}(f - f_0)|^2$ (where $\gamma = \text{scale factor} < 1$ and often $\ll 1$ and $f_0$ is a shift) from the spectral estimate one may be able to investigate low level structure in the estimate.

This approach was developed first in the 2-D context of radio interferometry and is the core of the algorithm ‘CLEAN’.

Roberts et al. 1987 (A.J.) have developed the CLEAN technique for 1-D spectral analysis of time series.
CLEAN

window \( S(t) \)

\[ \mathcal{F} \]

spectrum of window \( |\tilde{S}(f)|^2 \)

find largest amplitude in spectrum: \( A(f_{\text{max}}) \) at \( f=f_{\text{max}} \)

\[ \tilde{Z}(f) \times A(f_{\text{max}}) \] and shift \( f \rightarrow f_{\text{max}} \)

subtract \( g A(f_{\text{max}}) (\tilde{Z}(f-f_{\text{max}}))^2 \) from data spectrum

\[ g = \text{loop gain} \]

no

one residuum consistent with noise?

\[ j \leq 1 \]

\[ \sum c_j S(f) + \sum w_i(f) \]

CLEAN spectrum

\[ R(f) \]

\[ c_j \text{ = CLEAN component} \]

\[ w_i \text{ = residual spectrum} \]
II. Data Adaptive Spectral Analysis: The effective window function depends on the data and can vary with frequency.

Spectral estimators exist which have effective spectral windows that adapt to the properties of the signal (e.g. sampling, signal-to-noise, type of signal...); the effective window is produced integrally. As such, the window may be optimized with respect to the data, but the analyst to some extent has lost control over it and its properties (e.g. the effective number of degrees of freedom).

Data adaptive spectral estimators are generally nonlinear in the correlation function estimates.

We will consider the following data adaptive estimators:

1. Maximum entropy method.
2. “Maximum likelihood” (in quotes because it does not result from maximizing a likelihood function, but rather is closely related to calculating an ML estimate of the power at the output of a filter).
The philosophy of entropy maximization dictates that we make no assumptions about missing data.

We will consider the ME estimator in greatest detail and we will show that the ME spectral estimator is equivalent to

1. fitting an autoregressive (AR) model to the data and finding the spectrum of the AR model.

2. applying a linear, Wiener-Hopf prediction filter to the time series, thus extending the TS by some allowable amount, and finding the spectrum of the extended TS.

3. extending the correlation function according to an AR model.

In effect, ME techniques fill in gaps (interpolate) or extend the data (extrapolate) in a way that is consistent with the available data but also in such a way that entropy is maximized. This maximizes the uncertainty about the missing data; i.e. the least amount of “structure” is imposed on the missing data.

In some instances, one can “superresolve” features in the spectrum; this can be viewed as “beating” the uncertainty principle because the effective resolution in the frequency domain can be as large as

\[ \Delta f_{ME} \rightarrow 1/2T \]

\((T = \text{length of time series})\), i.e. twice the resolution of the conventional Bartlett type estimate.
This occurs (for some cases) because the MEM effectively predicts missing data from known data.

The regime in which this works best is where

$$T \sim \text{correlation time of the process}$$

because, at best, one can predict $\sim$ one correlation time into the future.

According to the ME philosophy, the sidelobes (Gibbs phenomenon) associated with spectral estimators based on Fourier transforms (i.e. either the smoothed $|F.T.|^2$ of the time series data or the F.T. of a correlation function) are a penalty that is consequence of “assuming” that unknown data are zero. (One is not so naive as to really assume the data are zero, but if one did, the same estimator would result.)
How can the data be extended?

We may have extensions of the form

I. Predict $x(t)$ forward by approximately one correlation time ($W_x$).

The new resolution will be $\delta f \sim T'^{-1} < T^{-1}$

II. Extend the corelation function, $R_x(\tau)$ or $C_x(\tau)$
Data adaptive methods help only when there really is *significant* missing data.

If $T = \text{actual time series duration} \gg W_x = \text{correlation time}$ then $T' \approx T + W_x \approx T \Rightarrow \text{no advantage.}$

But if $T \approx W_x$ so that $T' \sim 2T$ then a factor of 2 increase in resolution may result (super resolution)

*Example*: The case of 2 sinusoids: $x(t) = A \sin \omega_1 t + B \sin \omega_2 t + n(t)$

If $\delta f = T^{-1} \ll |\omega_2 - \omega_1|$ then F.T. methods will suffice in resolving the spectral lines.

However, if $T^{-1} \sim |\omega_2 - \omega_1|$, then ME estimates may provide better resolution.
Other Issues:

I. Bayesian approaches: we may want to ask a different question than “what is the power at a given frequency?” Rather, we may want to ask “what is the PDF for the frequency of a sinusoidal signal?”

II. Ad hoc approaches for fixing some aspects of Fourier based estimators (leakage effects for red power spectra).
Maximum Entropy Spectral Estimate

- The MESE is identical to fitting an autoregressive (AR) model to a time series

\[
X_t = \sum_{j=1}^{M} \alpha_j X_{t-j} + n_t
\]

\[
\hat{S}(f) = \frac{P_M}{\left|1 + \sum_{j=1}^{M} \alpha_j e^{2\pi i f j \Delta \tau}\right|^2}
\]

- The unknowns are \(P_M, \alpha_j, j=1,M\)
- We will show this equivalence in a later lecture
Maximum Entropy Spectra of Red Noise Processes

1. Generate a realization of red noise:
   • Spectral domain:
     – Generate complex white, Gaussian noise
     – Shape it according to \([S(f)]^{1/2}\)
     – Inverse FFT to time domain

2. Find the best fit autoregressive model by minimizing fitting error against the ‘order’ of the model
   • AR model fitting is equivalent to maximizing entropy

3. Find the Fourier spectrum of the AR model
$S(f) \propto f^{-0.0}$
$S(f) \propto f^{-1.0}$

$S(f) \propto f^{-2.0}$

$S(f) \propto f^{-3.0}$

$S(f) \propto f^{-4.0}$

$S(f) \propto f^{-5.0}$

$S(f) \propto f^{-6.0}$
Leakage effects for processes with "red" power spectra:

"red" $\Rightarrow$ more power at low frequency than high.

Recall leakage of a discrete spectral line:

\[ \hat{S}(f) = \left( \frac{\sin 2\pi ft}{2\pi ft} \right)^2, \quad T = \text{length of data set} \]

it is obvious that $\hat{S}(f) \propto f^{-2}$. (The envelope of oscillators)

For power law spectra $S(f) \propto f^{-n}$ with $n > 0$ there is excess power at low frequencies. As $n$ increases this excess increases dramatically, and leakage for it can contaminate all frequencies. The DFT-based spectral estimator based on:

\[ S(f) \]

\[ f^{-2} \]

\[ f \]

\[ S(f) \]

\[ f^{-2} \]

\[ \text{asymptotic slope} \Rightarrow 2 \]

\[ f \]
3. Computer Experiments

Using the program of McClellan et al. [1979], we designed filters whose squared magnitude had a power law response in the frequency range 0.01 to 0.5, with indices of 0.5 to 5 in steps of 0.5 (10 filters in all). The impulse response of the filter with index 2.0 is shown in Fig. 1(a). The impulse response is symmetric about the t = 0 point, and thus the frequency response is zero phase; that is, it is purely real. The frequency response is the Fourier transform of the impulse response, and it may be approximated as closely as desired by using an FFT of the impulse response augmented by a sufficiently large number of zeroes. The process of augmenting a discrete function with a large number of zeroes before performing a finite Fourier transform is called zero padding and results in a closer spacing of the transformed

\[ n = 2 \]

![Graphs showing impulse response, squared magnitude, frequency response, and input/output signals.]

**Figure 1.** (a) Impulse response of a filter designed to have a squared power law frequency response with a slope of -2. A total of 201 weights are used. (b) Squared frequency response of the filter. The slope is -2.0000. (c) The 1000-point sample of Gaussian white noise used as input to the filter. (d) The 800-point output; at each end, 100 points are lost because the symmetrical filter is 201 weights long.
values. Effectively, zero padding in the input produces interpolation in the output. Since the frequency response of a digital filter is the discrete Fourier transform of the (essentially discrete) impulse response, with zero padding the discrete Fourier transform can be approximated as closely as desired by the finite Fourier transform. Figure 1(b) shows the squared frequency response of the filter with index 2, approximated using a 2048-point FFT. The actual slope, obtained by fitting a straight line to the computed points using least squares, is $-2.0000$.

Using the filter described in Figs. 1(a) and (b), the sample of Gaussian white noise, shown in Fig. 1(c), produced the realization of a power law process with index 2 (the FIR filter output) shown in Fig. 1(d). The maximum entropy spectrum using five prediction error filter weights is given in Fig. 2(a), and the periodogram (unaveraged, point-by-point square of magnitude of FFT) is shown in Fig. 2(b). The MEM spectrum is much smoother than the FFT spectrum: there is much less variance from one frequency estimate to the next, and the shape itself is nearly linear, reflecting the linearity of the true spectrum. For each spectrum the slope is found by fitting a straight line to the power spectral density (PSD) estimates using least squares. In this procedure, all of the computed points shown in Fig. 2(a) or (b) are used to fit a straight line in the form

$$\log y = -m \log x + A,$$

where $m$ is the desired spectral index, using the method of least squares. The MEM slope is $-1.9578$, and the periodogram slope is $-1.9397$. Thus the smoothed behavior of the periodogram in this case is acceptable.

![Figure 2](image)

**Figure 2.** (a) Maximum entropy spectrum of the signal from Fig. 1(d). The ordinate is 10 log$_{10}$ (PSD). (b) Periodogram of the same signal.

When the spectral index is increased to 4, the resulting impulse response, squared frequency response, white noise input, and red noise realization are those given in Fig. 3. The MEM spectrum, this time using 10 weights, and the periodogram are given in Fig. 4.
Figure 3. (a) Impulse response of a filter designed to have a squared power law frequency response with a slope of $-4$. A total of 301 weights are used. (b) Frequency response of the filter. The slope is $-4.0009$. (c) The 1000-point sample of Gaussian white noise used as input to the filter. (d) The 700-point output. At each end, 150 points are lost because the symmetrical filter is 301 weights long.

Figure 4. (a) Maximum entropy spectrum of the signal from Fig. 3(d). The ordinate is $10 \log_{10} (\text{PSD})$. (b) Periodogram of the same signal.
ME spectra of power law processes

Figures from Fougere in Smith & Erickson.

"End-matching" is the procedure of
1. fitting a straight line to the 1st and last data points
2. subtracting line from all data points in time series

This "end hockey" is aimed at reducing the discontinuity in the periodically extended data between beginning and end.

```
---+---+---+
|    |    |    |
```

```
| time series | periodic expansion | end matched |
```

"Windowing" de-emphasizes the beginning and end of the time series ⇒ no problem with discontinuities (or less of a problem, anyway)
Figure 5. Observed FFT (periodogram) index versus the MEM index (index is the negative of the slope). There are 100 independent realizations of the power law process for each of the 10 indices 0.5, 1.0, ..., 5.0. In every case the same time series was used as input to MEM and to FFT (periodogram). (a) Raw data. (b) End-matched data. (c) Windowed data. (d) End-matched and windowed data.