Lecture 11
- Red spectra, end-matching examples (quick)
- Maximum entropy method
- Information and Entropy

Assignment 2: Due 21 March

Reading:
- A Bayesian Approach to Spectral Analysis (pdf file under Lecture 9)

Code:
- AR processes (see definitions in "useful processes" document on web page
  (under Lecture 5)
  - Jupyter notebook on AR time series example
  - Run by changing the a1, a2 parameters for an M=2 model
  - Experiment with other values of M
  - Add plotting of autocorrelation function

References:
- "Spectral Analysis of Signals" Stoica & Moses (427 pp)
  - http://user.it.uu.se/~ps/SAS-new.pdf
- "Bayesian Spectrum Analysis and Parameter Estimation" Brethorst (220 pp)

Webpage: www.astro.cornell.edu/~cordes/A6523

SPECTRAL ANALYSIS OF SIGNALS

Petre Stoica and Randolph Moses

http://user.it.uu.se/~ps/SAS-new.pdf

Contents
- Basic Concepts
- Nonparametric Methods
- Parametric Methods for Rational Spectra
- Parametric Methods for Line Spectra
- Filter Bank Methods
- Spatial Methods
- Appendices
  - Linear Algebra and Matrix Analysis Tools
  - Cramer-Rao Bound Tools
  - Model Order Selection Tools
Testing isotropy in the Two Micron All-Sky redshift survey with information entropy

Biswajit Pandey

Department of Physics, Visva-Bharati University, Santiniketan, Birbhum, 713103, India

ABSTRACT

We use information entropy to test the isotropy in the nearby galaxy distribution mapped by the Two Micron All-Sky redshift survey (2MRS). We find that the galaxy distribution is highly anisotropic on small scales. The radial anisotropy gradually decreases with increasing length scales and the observed anisotropy is consistent with that expected for an isotropic Poisson distribution beyond a length scale of 90 h^{-1}Mpc. Using mock catalogues from N-body simulations, we find that the galaxy distribution in the 2MRS exhibits a degree of anisotropy compatible with that of the CDM model after accounting for the clustering bias of the 2MRS galaxies. We also quantify the polar and azimuthal anisotropies and identify two directions (l,b) = (150°, -15°), (l,b) = (140°, -15°) which are significantly anisotropic compared to the other directions in the sky. We suggest that their preferential orientations on the sky may indicate a possible alignment of the Local Group with two nearby large scale structures. Despite the differences in the degree of anisotropy on small scales, we find that the galaxy distributions in both the 2MRS and the CDM model are isotropic on a scale of 90 h^{-1}Mpc.

Key words: methods: numerical - galaxies: statistics - cosmology: theory - large scale structure of the Universe.
Red Noise: Challenges for Spectral Estimation

A Bayesian test for periodic signals in red noise

S. Vaughan*  
X-Ray and Observational Astronomy Group, University of Leicester, Leicester, LE1 7RH, U.K.

Accepted 2009 October 14. Received 2009 October 14; in original form 2009 September 11

ABSTRACT

Many astrophysical sources, especially compact accreting sources, show strong, random brightness fluctuations with broad power spectra in addition to periodic or quasi-periodic oscillations (QPOs) that have narrower spectra. The random nature of the dominant source of variance greatly complicates the process of searching for possible weak periodic signals. We have addressed this problem using the tools of Bayesian statistics; in particular using Markov chain Monte Carlo techniques to approximate the posterior distribution of model parameters, and posterior predictive model checking to assess model fits and search for periodogram outliers that may represent periodic signals. The methods developed are applied to two example datasets, both long XMM-Newton observations of highly variable Seyfert 1 galaxies: RE J1034 + 396 and Mrk 766. In both cases a bend (or break) in the power spectrum is evident. In the case of RE J1034 + 396 the previously reported QPO is found but with somewhat weaker statistical significance than reported in previous analyses. The difference is due partly to the improved continuum modelling, better treatment of nuisance parameters, and partly to different data selection methods.

Key words: Methods: statistical – Methods: data analysis – X-rays: general – Galaxies: Seyfert

Presents a nice summary of Bayesian methodology + comparison with periodograms.

Applies to time series with red noise + periodicity but does not deal with spectral leakage for steep spectra.
The impact of red noise in radial velocity planet searches: Only three planets orbiting GJ581?

Roman V. Baluev

Central (Pulkovo) Astronomical Observatory of Russian Academy of Sciences, Pulkovskoe sh. 65/1, St Petersburg 196140, Russia

Abstract
We perform a detailed analysis of the latest HARPS and Keck radial velocity data for the planet-hosting red dwarf GJ581, which attracted a lot of attention in recent time. We show that these data contain important correlated noise component (‘red noise’) with the correlation timescale of the order of 10 days. This red noise imposes a lot of misleading effects while we work in the traditional white-noise model. To eliminate these misleading effects, we propose a maximum-likelihood algorithm equipped by an extended model of the noise structure. We treat the red noise as a Gaussian random process with exponentially decaying correlation function.

Using this method we prove that: (i) planets b and c do exist in this system, since they can be independently detected in the HARPS and Keck data, and regardless of the assumed noise models; (ii) planet e can also be confirmed independently by the both datasets, although to unveil it in the Keck data it is necessary to take the red noise into account; (iii) the recently announced putative planets f and g are likely just illusions of the red noise; (iv) the reality of the planet candidate GJ581 d is questionable, because it cannot be detected from the Keck data, and its statistical significance in the HARPS data (as well as in the combined dataset) drops to a marginal level of \( \sim 2 \sigma \), when the red noise is taken into account.

Therefore, the current data for GJ581 do not support existence of more than four (or maybe even only three) orbiting exoplanets. The planet candidate GJ581 d requires serious observational verification.

Key words: planetary systems - stars: individual: GJ581 - techniques: radial velocities - methods: data analysis - methods: statistical - surveys

Simulating power-law noise

Applications: many phenomena in nature are processes with spectra that are power-law in form (temporally or spatially)

\[ S(f) \propto f^{-\alpha} \quad f_0 \leq f \leq f_1 \]

Can use a linear filter with impulse response \( h(t) \):

\[
\begin{align*}
   x(t) & \rightarrow h(t) \rightarrow y(t) \\
   y(t) &= x(t) * h(t) \\
   \tilde{Y}(f) &= \tilde{X}(f) \cdot \tilde{H}(f) \}
\end{align*}
\]

\[ (|\tilde{Y}(f)|^2) = (|\tilde{X}(f)|^2)|\tilde{H}(f)|^2 \]

Let \( x(t) \) = realization of white noise and \( H(f) = \text{sqrt(shape)} \) of spectrum that is wanted
Procedure

- Generate spectrum in frequency domain
- Generate white noise realizations for real and imaginary parts of \( H(f) \)
  - random number generator: gaussian, uniform
- Multiply by \( f^{-\alpha/2} \text{ for } f_0 \leq f \leq f_1 \)
- Fill vector to force Hermiticity
- Do inverse DFT to get real time-domain realization of the noise
- Can be applied to any shape of spectrum of course
- What will the statistics be in the time domain?

\[ f^3 \text{ spectrum} \]
Spectral index = 3.0 before end matching

Spectral index = 3.0 with slope lines

Slope = -2
Slope = -3
Underlying Issue
The time series is periodic \(\rightarrow\) discontinuities

Data
Spectral index = 3.0 with slope lines

Fixes
• Last lecture we discussed window functions as a way to lessen sidelobe effects
  – Tapering of the time series
  – E.g. a Gaussian function centered on the middle of the time series will make the time series and its derivatives continuous (to the extent that the Gaussian has fallen off at the edges of \([0, T]\)
• Here is an *ad hoc* fix called “end matching” which matches the two endpoints of the time series
Ad hoc fix = “End Matching”

Spectral index = 3.0 with slope lines

Subtract line that connects the end points.
This makes the periodically extended time series continuous

Data

Spectral index = 3.0 end-match line

Subtract line that connects the end points.
This makes the periodically extended time series continuous

Data

Spectrum = FFT

Spectral index = 3.0 after end matching

$f^5$ spectrum
Data
Spectral index = 5.0 before end matching

Spectrum = \[ |FFT|^2 \]

Data
Spectral index = 5.0 with slope lines

Spectrum = \[ |FFT|^2 \]
Slope = -5
Slope = -2

→ End matching not working so well
f^6 spectrum

Spectral index = 6.0 before end matching
Spectral index = 6.0 with slope lines

Spectrum = \[ FFT \]

Slope = -2

Spectrum = \[ FFT \]^2

Slope = -6
Why is end-matching failing?

- Clue: end-matching is progressively worse for steeper spectra
Why is end-matching failing?

- Clue: end-matching is progressively worse for steeper spectra
- Answer:

Though end matching makes the time series continuous (no discontinuities) it does not prevent discontinuities in the derivatives of the time series. As the spectrum gets steeper, the higher order derivatives become more important.
Figure 2. (a) Maximum entropy spectrum of the signal from Fig. 1(d). The ordinate is 10 log₁₀ (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 NEM 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 -4.0008. (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 log₁₀ (PSD). (b) Periodogram of the same signal.
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
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.
Maximum Entropy Methods

• Probabilities \( H = - \sum p_i \log p_i \)
  – Combine with constraint equations to determine \( p_i \)
  – No constraints ➔ flat distribution

• Spectra (and images) \( H = \text{integral of } \log S(f) \)
  – Constraints are from measurements
  – The spectrum is obtained by maximizing entropy subject to the constraints (what constraints do we have?)
  – For a Gaussian process, the resulting expression is the same as optimally fitting an AR model to a time series (proven in a paper cited in a later slide)
  – Why Gaussian?

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 ⇒ 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.

\[ P\{\text{oil shortage } | t < 1973\} \ll 1 \]
\[ P\{\text{oil shortage } | t = 2011\} \ll 1. \]

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. \( \lim_{P \to 0^+} f(P) = 0 \)
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, 
   \[ I_{12} = f(P_{12}) = f(P_1 P_2) = f(P_1) + f(P_2) \]
The only functions that satisfy these conditions are logarithmic ones
   \[ f(P) = -\log_b P \]
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 \( \zeta \rightarrow I(\zeta) = -\log P \)).

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:

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

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

\[ H_{\text{max}} = \log 2 = - \log \left( \frac{1}{2} \right) = 1 \]

\[ \text{n events: } H \text{ is maximized when all events are equiprobable, if there are no other constraints:} \]

\[ P_j = \frac{1}{n}, \quad j = 1, n \]

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

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

\[ \frac{\partial J}{\partial P_i} = \frac{\partial H}{\partial P_i} + \lambda \sum_j \frac{\partial P_j}{\partial P_i} \]

\[ = - \sum_j \left[ \frac{\partial P_j}{\partial P_i} \log P_j + P_j \frac{\partial \log P_j}{\partial P_i} - \lambda \frac{\partial P_j}{\partial P_i} \right] \]

\[ = - \sum_j \{ \delta_{ij} \log P_i + \delta_{ij} - \lambda \delta_{ij} \} \]

\[ = - \log P_i - 1 + \lambda = 0 \]

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)

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

and \( \log P_i = \lambda - 1 \Rightarrow \lambda = 1 + \log \frac{1}{n} \)

\[ \lambda = 1 - \log n \]
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 Definitions for Continuous Case and coordinate transformations

- Slides commented out for lecture but will appear in PDF file on web page

**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,

\[ H_X = - \int d\mathbf{x} f_X(\mathbf{x}) \log f_X(\mathbf{x}) \]

\( H \) 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
\[ \mathbf{y} = A \mathbf{x} \quad \text{with} \quad y_i = \sum_{j=1}^{n} a_{ij} x_j \]
which has a Jacobian
\[ J = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \ldots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \ldots & \frac{\partial y_n}{\partial x_n} \end{vmatrix} = |a_{ij}|^{-1} \]

Let \( H_Y = -\int dy f_Y(y) \log f_Y(y) \]
\[ = H_X - (\log J) \]
\[ = H_X - (\log |a_{ij}|)^{-1} \]
\[ = H_X + \log |a_{ij}| \]

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
\[ f_Y(y) = f_X(x(y)) \left| \frac{dy}{dx} \right| \]
if \( y = g(x) \) is a single valued function. Therefore
\[ H_Y = -\int dy f_Y(y) \log f_Y(y) - \int dy f_X(x(y)) \left[ \log f_X(x) - \log \left| \frac{dy}{dx} \right| \right] \]

Now convert back to a integration over \( x \)
\[ y = g(x), \quad dy = \frac{dy}{dx} \left| \frac{dy}{dx} \right| dx \]
\[ H_Y = -\int dx f_X(x) \log f_X(x) - \log \left| \frac{dy}{dx} \right| \]
\[ = -\int dx f_X(x) \log f_X(x) + \int dx f_X(x) \log \left| \frac{dy}{dx} \right| \]
\[ \Rightarrow H_Y = H_X + \left( \log \left| \frac{dy}{dx} \right| \right) \]

\[ H_Y = H_X + \left( \log \left| \frac{dy}{dx} \right| \right) \]
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 $H = - \int f(x) \log f(x)$ 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 $f(x)$:

$$
\delta J = \int dx \delta f(x) \left( -\log f(x) + \lambda x^2 + \mu \right) - f(x) \frac{\delta f(x)}{f(x)} = 0
$$

Class: you should show that this procedure yields a maximum by considering the second derivative of $J$ and show that $\frac{\partial^2 J}{\partial f^2} < 0$. 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:

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

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

Recall Boltzmann’s $H$ theorem:

$$
\frac{1}{2} \sigma^2 = E = \text{energy} \text{ with } \sigma^2 = KT.
$$

**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 $k$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, $H = 0 \Rightarrow$ 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.
Maximum Entropy: General Solution for PDF with Constraints

First a simple case:

The entropy for a discrete random variable \( X \), which takes on values \( \{x_k, k = 0, \ldots, N-1\} \) with associated probabilities \( p_k \) is

\[
H = -\sum_k p_k \ln p_k.
\]

When there is a constraint on the \( n^{th} \) moment \( \langle x^n \rangle \) the PDF is found by maximizing the quantity

\[
J = H - \lambda_0 \sum_k p_k - \lambda_1 \sum_k x_k^n p_k
\]

which uses a Lagrange multiplier \( \lambda_0 \) for the normalization constraint and \( \lambda_1 \) for the constraint on the moment. This construction allows us to calculate the total variation with respect to the \( p_k \).

Taking increments \( \delta p_k \) we obtain

\[
\delta H = -\sum_k (\delta p_k \ln p_k + \delta p_k)
\]

\[
\delta J = \delta H - \lambda_0 \sum_k \delta p_k - \lambda_1 \sum_k x_k^n \delta p_k
\]

\[
= -\sum_k \delta p_k (\ln p_k + 1 + \lambda_0 + \lambda_1 x_k^n)
\]

\[
= 0.
\]

The factor in parentheses needs to vanish since \( \delta p_k \) is arbitrary (but small), so

\[
p_k = e^{-(\lambda_0+1)x_k^n} e^{-\lambda_1 x_k^n}
\]

\[ k = 0, \ldots, N - 1.\]

We can solve for the Lagrange multipliers by substitution into the constraint equations:

\[
\sum_k p_k = e^{-(\lambda_0+1)} \sum_k e^{-\lambda_1 x_k^n} = 1 \quad \Rightarrow \quad e^{-(\lambda_0+1)} = \sum_k e^{-\lambda_1 x_k^n}
\]

and

\[
p_k = \frac{e^{-\lambda_1 x_k^n}}{\sum_k e^{-\lambda_1 x_k^n}}.
\]

The Lagrange multiplier \( \lambda_1 \) needs to be obtained from the expression for the \( n^{th} \) moment,

\[
\langle x^n \rangle = \sum_k x_k^n \delta p_k = \frac{\sum_k x_k^n e^{-\lambda_1 x_k^n}}{\sum_k e^{-\lambda_1 x_k^n}}
\]

\( n = 1 \): Constraint on the mean and the \( x_k \) are constrained to be positive:

\[
\langle x \rangle = \sum_k x_k \delta p_k = \frac{\sum_k x_k e^{-\lambda_1 x_k}}{\sum_k e^{-\lambda_1 x_k}}
\]

By inspection, we see that this is an exponential distribution so \( \lambda_1 \) has to be the mean, \( \langle x \rangle \).

We can show this formally for the case where the \( x_k \) are uniformly spaced, \( x_k = k \delta x, \ k = 0, \ldots, N - 1 \):
Let $a = e^{-\lambda x}$. Then using a trick to get the numerator,

\[
\frac{1}{\delta x} \sum_{k} x_k e^{-\lambda x} = \sum_{k} k a^k = a \times \frac{d}{da} \sum_{k=0}^{N-1} a^k = a \times \frac{d}{da} \left( \frac{1 - a^N}{1 - a} \right) = a \times \frac{(1 - a)(-N a^{N-1}) - (1 - a^N)(-1)}{(1 - a)^2},
\]

we obtain

\[
\frac{\langle x \rangle}{\delta x} = \frac{1}{1 - a} - \frac{(1 - a) N a^{N-1}}{1 - a^N}.
\]

If $N \to \infty$ only the leading term matters and we can solve for $a$,

\[
a = e^{-\lambda x} = \frac{\langle x \rangle \delta x - 1}{\langle x \rangle / \delta x}
\]

A similar approach can be taken for a case where the variance is fixed. In the continuous limit, this yields a Gaussian PDF.

---

**General case with arbitrary constraints on moments**

(From Jaynes, *Probability Theory: The Logic of Science*, pp. 355-358)

Discrete RV: $x_n, n = 0, \ldots, N - 1$

Constraints on arbitrary functions of $x$: $f_k(x), k = 1, \ldots, M$.

The constrain equations are the expectations:

\[
\langle f_k(x) \rangle = F_k = \sum_{n=0}^{N-1} p_n f_k(x_n).
\]

Maximize entropy subject to constraints:

\[
J = H - \lambda_0 \sum_{n} p_n - \sum_{k} \lambda_k \langle f_k(x) \rangle
\]

\[
= -\sum_{n} p_n \left[ \ln p_n + \lambda_0 + \sum_{k} \lambda_k f_k(x_n) \right]
\]

Setting $\delta J = 0$ for arbitrary $\delta p_n$ and using normalization to unity to obtain $\lambda_0$ we get

\[
p_n = \frac{1}{Z(\lambda)} \sum_{k} \lambda_k f_k(x_n)
\]

where the partition function is

\[
Z(\lambda) = \sum_{\{n\}} e^{-\sum_{n} \lambda_k f_k(x_n)}
\]

Solving for the Lagrange multipliers clearly is nontrivial in general.
Maximum Entropy: Power Spectrum (short approach)

So far we know how to calculate the entropy of a random variable in terms of its PDF. For a univariate Gaussian PDF we have

\[ f_X(x) = (2\pi\sigma^2)^{-1/2} e^{-x^2/2\sigma^2} \]

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

\[ = \left\{ \frac{1}{2} \ln(2\pi\sigma^2) + \frac{X^2}{2\sigma^2} \right\} \]

\[ = \frac{1}{2} \left[ \ln(2\pi\sigma^2) + 1 \right] \]

\[ = \frac{1}{2} \left[ \ln(2\pi\sigma^2) \right] \]

When we maximize the entropy subject to constraints (from data), we only care about terms in the entropy that depend on relevant parameters. Here the only parameter is \( \sigma \) so the constant term does not matter. Notice that larger \( \sigma \) implies larger entropy, as we would expect for a measure of uncertainty.

When we maximize entropy, we may as well write it only in terms of the variance,

\[ H \approx \ln \sigma^2 + \text{constant}. \]

---

Heuristic “derivation” of the entropy rate expression:

Another way of viewing this is as follows. In calculating a power spectrum we are concerned with a second-order moment, by definition. Consequently, we can assume that the random process under consideration is Gaussian because:

1. we are maximizing the entropy (subject to constraints) and
2. given the second moment, the process with largest entropy is a Gaussian random process

Note that while this assumption is satisfactory for estimating the power spectrum (a second moment), it is not necessarily accurate when we consider the estimation errors of the spectral estimate, which depend on fourth-order statistics. If the central limit theorem can be invoked then, of course, the Gaussian assumption becomes a good one once again.

Imagine that the process under study is constructed by passing white noise through a linear filter whose system function is \( \sqrt{S(f)} \)

\[ n(t) \rightarrow \sqrt{S_x(f)} \rightarrow x(t) \]

Consequently, the Fourier transforms are related as

\[ \tilde{N}(f) \sqrt{S_x(f)} = \tilde{X}(f) \]

Now \( \tilde{N}(f) \) itself is a Gaussian random variable because it is the sum of GRV’s. Therefore, \( \tilde{X}(f) \) is a GRV and viewing it as a 1-D random variable we have that the entropy is

\[ H(f) = \frac{1}{2} \ln [2\pi e \sigma^2(f)] \]

but