AMERICAN UNIVERSITY HONORS CAPSTONE

Introduction to Stochastic Process Analysis

Spectral and Bayesian Techniques for Estimation and Prediction

Adam Ma'ruf

5/4/2010

This paper introduces two broad methods of analyzing stochastic processes. Stochastic processes have enormous potential for mathematical modeling across a wide variety of scientific disciplines, including physics, biology, and economics. The two methods discussed in this paper are Spectral Analysis—which deals with the detection of cyclic behavior—and Bayesian estimation—which provides a robust framework for predicting stochastic processes via Monte Carlo Simulation. MATLAB simulations are used throughout to guide the reader.

Part I: Spectral Estimation

Fourier Foundations:

To provide the best explanation of the transform, we start with a wholesome definition of Fourier series. We are going to approach Fourier series as an application of an inner product in the vector space of continuous functions. As Lay explains, the inner product space of all continuous functions on an interval $a \leq t \leq b$ is probably the most widely used inner product space for applications (Lay 433). We start by defining a polynomial that exists in $\mathrm{P_n}$. The "length" for p can be computed in the interval [a,b] by evaluating p at n+1 points of length Δt in [a,b]. The inner product of two polynomials p and q is:

$$\langle p, q \rangle = p(t_0)q(t_0) + \cdots p(t_n)q(t_n) = \sum_{i=1}^{n} p(t_i)q(t_i)$$

As it satisfies the inner product axioms (Lay 430-435). For large n, the inner product will be large, so we scale it down by diving by n+1. Now note that $\frac{1}{n+1} = \frac{\Delta t}{b-a}$, as both represent one discrete piece of the entire interval. Thus we write:

$$\frac{1}{n+1}\sum_{i=1}^{n} p(t_i)q(t_i) = \frac{\Delta t}{b-a}\sum_{i=1}^{n} p(t_i)q(t_i) = \frac{1}{b-a}\sum_{i=1}^{n} p(t_i)q(t_i)\Delta t$$

Obviously, the final representation is a Riemann sum that converges to an integral as $\Delta t \to 0$:

$$\frac{1}{b-a}\int_a^b p(t)q(t)dt$$

Where the scalar is not essential, and is often omitted (Lay 439).

Putting polynomials in C[a,b] aside, let us consider a function in $C[0,2\pi]$. A continuous function can be approximated by linear combinations of sine and cosine functions, otherwise known as trigonometric polynomial:

$$\frac{a_0}{2} + a_1 \cos 1t \ldots a_n \cos nt + b_1 \sin 1t \ldots b_n \sin nt$$

We are going to define an inconsistent linear system $Ax = f$, in other words an approximation of the function, where the A is defined as:

$$A = \begin{bmatrix} \frac{a_0}{2} & a_1 \cos 1t & b_1 \sin 1t \\ \vdots & \vdots & \vdots \\ \frac{a_0}{2} & a_n \cos nt & b_n \sin nt \end{bmatrix}$$

Note that the 2nd and 3rd columns of a are orthogonal with respect to the inner product $C[0,2\pi]$. The function f may not exist in the column space of A, but the best approximation would be the function's orthogonal projection onto the column space. This minimizes the distance between the function evaluated at t and the trigonometric polynomial. This approximation is called the Fourier approximation to f on $[0,2\pi]$. The orthogonal projection shows that the coefficients $a_k, k = 1, \dots n$ and $b_k, k = 1, \dots n$ are:

$$a_k = \frac{\langle f, \sin kt \rangle}{\langle \sin kt, \sin kt \rangle} \, and \; b_k = \frac{\langle f, \cos kt \rangle}{\langle \cos kt, \cos kt \rangle}$$

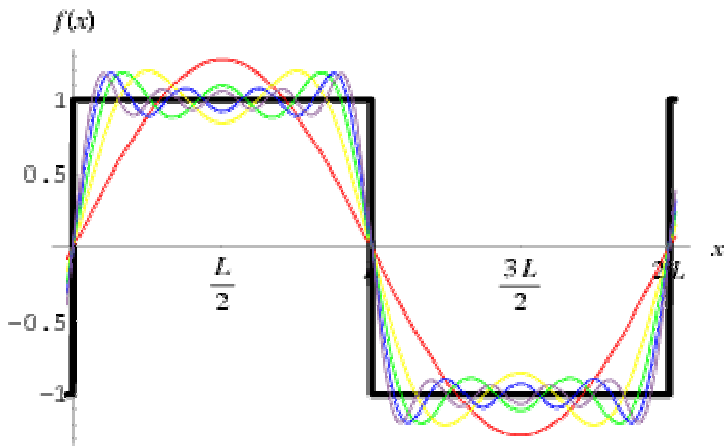
A simple computation shows that $\langle \sin kt, \sin kt \rangle = \pi$ and $\langle \cos kt, \cos kt \rangle = \pi$. So we are left with

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin(kt)\, dt \;\; and \; b_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos(kt)\, dt$$

Having found the value of the coefficients that minimize distance between function and polynomial, the best approximation to the function in terms of sines and cosines has been approximated, and can be written concisely as:

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos kt + b_k \sin kt$$

This is known as the Fourier series for function f. The term $a_k \sin kt$ is, for example, the projection of f onto the one-dimensional subspace spanned by $\sin kt$. As n increases, the distance between function and trigonometric polynomial approaches zero. The coefficient in this equation is usually dropped in the general form of the Fourier series.



A Fourier series approximating a step-wise function, from Wolfram (wolfram.com)

Fourier Analysis: The Transform

The Fourier series can be extended to complex coefficients, and such an extension is what leads to the extraordinary Fourier Transform. If we instead right the Fourier series with both sines and cosines multiplied by the same coefficient a_k, but the sines are also multiplied by i, the imaginary number, then we can right the series as:

$$f(t) = \sum_{k=1}^{n} a_k \cos kt + i a_k \sin kt$$

Because the two terms still represent orthogonal columns of a matrix representing a trigonometric polynomial. Now, we take advantage of Euler's phenomenal identity:

$$e^{i\theta} = \cos\theta + i \sin\theta$$

This allows us to write the Fourier series as:

$$f(t) = \sum_{k=1}^{n} a_k e^{ikx}$$

Now let us take the integral of $f(t)$, which is periodic in $[-\pi, \pi]$, multiplied by another complex variable, e^{-imx}. This integral represents the "amount" of the wave represented by e^{imx} that is represented by the foureier series of $f(t)$ (Wikipedia, Fourier Transform).

$$\int_{-\pi}^{\pi} f(t)\, e^{-imx} dx = \int_{-\pi}^{\pi} \sum_{k=-\infty}^{\infty} a_k e^{ikx}\, e^{-imx} dx = \sum_{k=-\infty}^{\infty} a_k \int_{-\infty}^{\infty} e^{i(k-m)x} dx$$

$$= \sum_{k=-\infty}^{\infty} a_k \int_{-\infty}^{\infty} \cos(k-m)x + i \sin(k-m)x\, dx = 2\pi a_m, (Wolfram)$$

Notice how the multiplication of the function's Fourier series and the wave represents a difference between the Fourier series and the wave: the integral thus represents the total amount of the wave that is in the Fourier series of the function. The above derivation is made to prove that the integral is equal to $2\pi a_m$. Then, it is obvious that:

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\, e^{-ikx} dx$$

We now have functional representations of $f(t)$ and a_k for the function periodic in $[-\pi, \pi]$. We can generalize this to a function periodic in $[-\frac{L}{2}, \frac{L}{2}]$:

$$f(t) = \sum_{k=-\infty}^{\infty} a_k e^{i2\pi kx/L} \ and\ a_k = \frac{1}{L}\int_{-L/2}^{L/2} f(t)\, e^{-i2\pi kx/L} dx$$

Note that our constant, a_k, can be considered a function of $\frac{k}{L}$. Written as $a_k = \tilde{f}(\frac{k}{L})$. Now, here is the ingenious part of the derivation. For simplicity sake, write $\varepsilon = \frac{k}{L}$. Then, $\Delta\varepsilon = \frac{(k+1)}{L} - \frac{k}{L} = \frac{1}{L}$. Then the periodic function $f(t)$ can be written as:

$$f(t) = \sum_{k=-\infty}^{\infty} \tilde{f}(\varepsilon) \times e^{i2\pi x\varepsilon}\Delta\varepsilon$$

Extraordinarily, this is a Riemann sum which in the limit becomes:

$$\lim_{L\to\infty} f(t) = \lim_{L\to\infty} \sum_{k=-\infty}^{\infty} \tilde{f}(\varepsilon) \times e^{i2\pi x\varepsilon}\Delta\varepsilon = \int_{-\infty}^{\infty} \tilde{f}(\varepsilon) e^{i2\pi x\varepsilon} d\varepsilon$$

This is the definition of the inverse Fourier Transform, ie, constructing a function in the time domain out of a function of frequencies—cycles per unit of time. To construct a function of frequencies from a function of time, the inverse of the above function is used: this is known as the Fourier Transform:

$$\tilde{f}(\varepsilon) = \int_{-\infty}^{\infty} f(t) e^{-i2\pi x\varepsilon} d\varepsilon$$

Understanding the Fourier Transform:

We have gone through great lengths to derive the Fourier Transform because it plays such an important role in classical signal processing techniques. Before moving on to explain how the transform can be applied to stochastic processes, we want to augment the mathematical development with a more intuitive understanding.

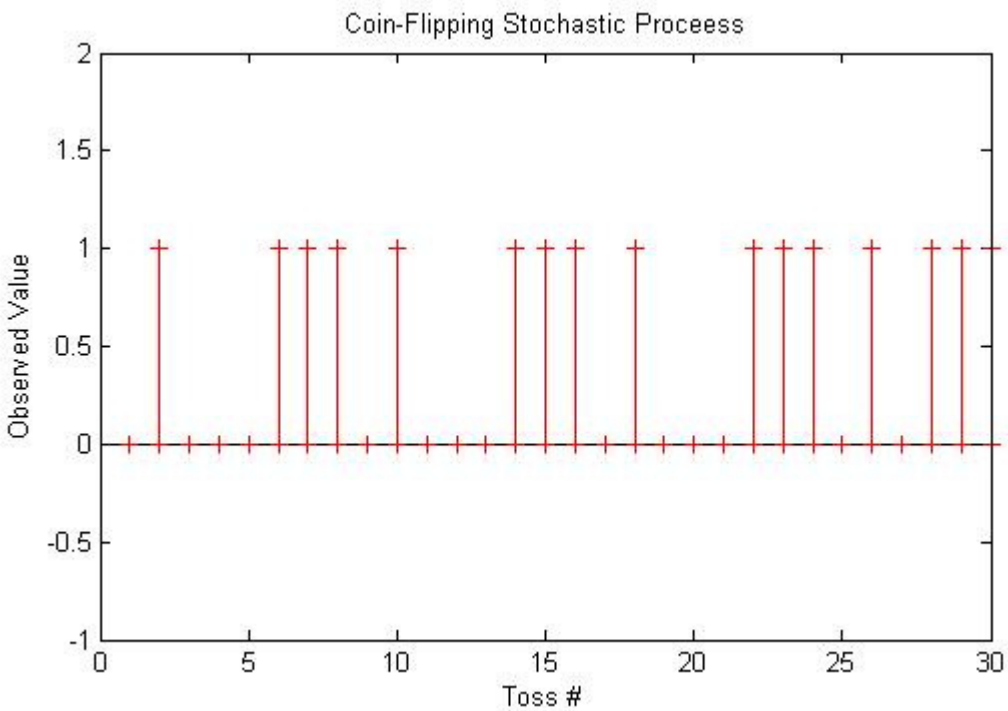
A function of $\frac{k}{L}$ is a function of frequency (remember kt is the angular frequency of the sinusoidal waves making up our initial trigonometric polynomial. So, the Fourier transform is

transforming our function of time into a function of frequency, or phrased differently, a function of how many occurrences in a unit interval of time. Theoretically ,this is all the values of k for which the integral is evaluated. Thus, the transform measures how much of each frequency is present in our time-valued function. Later on, we will provide a graphical illustration with the transform of time-valued stochastic processes. But first, the basics of stochastic processes must be explained.

Stochastic Processes: Introduction

A precise mathematical definition of a stochastic process is a random vector of infinite dimension. However, this definition does not motivate further studies of stochastic processes, when the motivation should be great: stochastic processes have enormous potential for modeling phenomena across the hard and social sciences.

Consider a coin toss. The result of the toss is obviously random, and if X represents the face of the coin after the toss (X = 1 if heads and X = 0 if tails), then X is a textbook example of a random variable. Now suppose that every 5 seconds, you toss the same coin. You do this indefinitely, in essence creating a string of random variables that are indexed in time. This is a basic example of a stochastic process: a process which takes different values at different time periods, and each value at each time is the outcome of a random experiment.



A simple stochastic process; note that there is no obvious pattern in which value it takes.

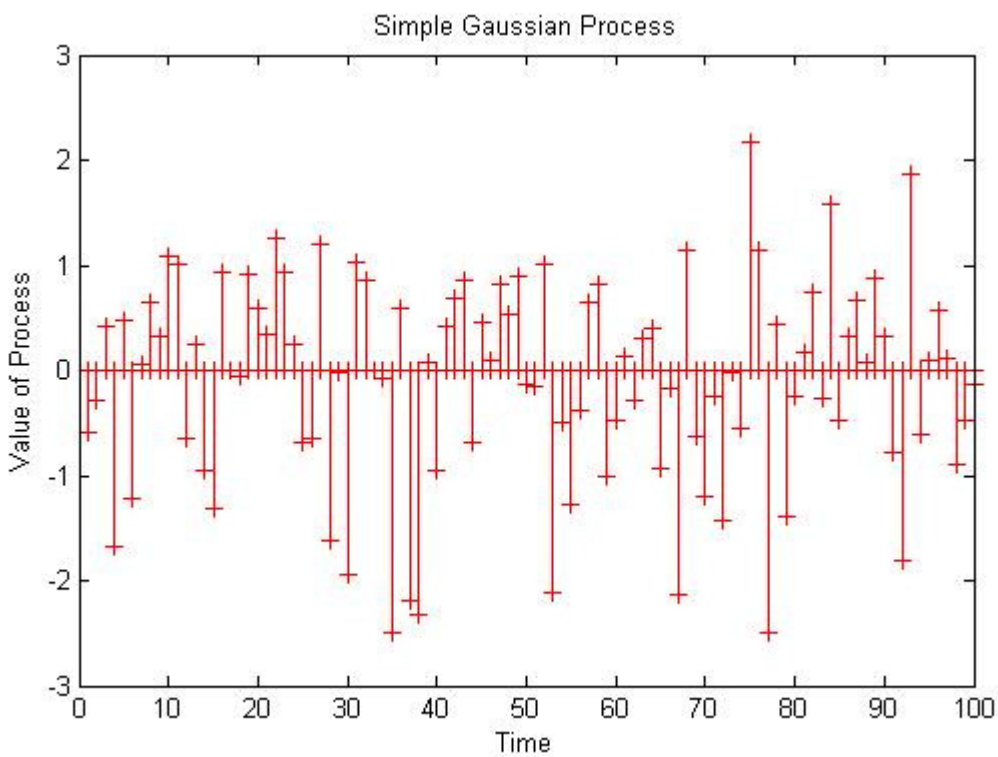
So, a stochastic process is simply a process whose value at any given time follows a probability distribution.

As Kay says, “There are several types of random processes that have found wide application because of their realistic physical mdeling yet relative mathematical simplicity” (Kay, 674). One of the processes he is speaking of is the Gaussian process.

Gaussian Processes have many very appealing properties, owing to the fact that they are created from the very appealing Gaussian distribution. For instance, the joint PDF of any set of samples is a multivariate Gaussian distribution. If **X** is a Gaussian random process, then $\boldsymbol{X} = [X_1, X_2, \dots, X_N]^T$ and its pdf is the joing pdf of the random variables $X_1, X_2, \dots, X_N$:

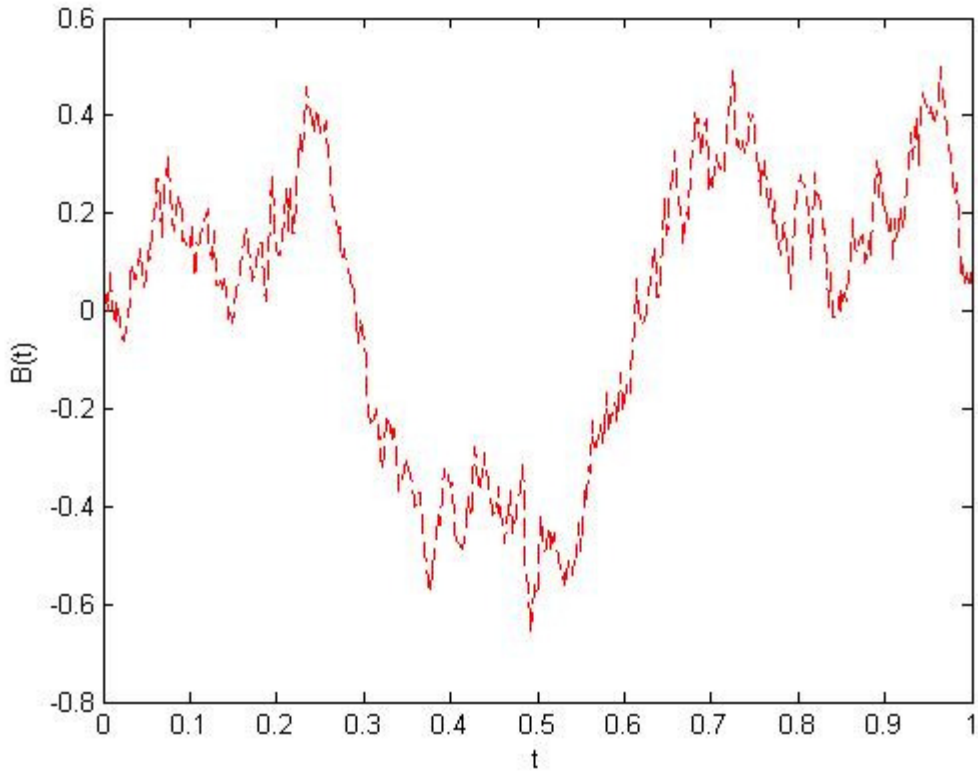
$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2} \times det^{\frac{1}{2}}(\boldsymbol{C})} \times exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Where $\boldsymbol{\mu}$ is the vector of means, and **C** is the covariance matrix. If the Gaussian random variables are independent, then C is just the vector of variances. Additionally, the first two moments completely define it (Kay, 674). Also, a linear transformation of a Gaussian distribution is still Gaussian, so we may apply linear transformations on a Gaussian process and still have a Gaussian process as a result.



Simulation of a Gaussian process comprised of independent standard normal Gaussian distributions.

Of course, stochastic processes can get much more complicated. For instance, consider Brownian motion, discovered by Robert Brown in 1827 and used to describe the minute changes in position of microscopic pollen particles (Wiersema). The movement of Brownian motion is continuous and each increment in its position follows a normal distribution with mean 0 and variance proportional to the time step ‘t’:



A single path of Brownian motion. Note that simulating the code for this graph again would produce a different path

An even more complicated stochastic process is based on a differential equation with one rate of change described with respect to Brownian motion. The solution to that stochastic differential equation is graphed below, and is used to model the movement of stock prices.

Clearly, the breadth of stochastic processes and the phenomena they can model is vast. Before moving on to Fourier Analysis of these processes, we are going to introduce two very important concepts: stationarity and autocorrelation.

Stationarity and Autocorrelation

One important caveat of modeling with stochastic processes is that much analysis can only be done on processes whose moments do not vary across time; put differently, the value of the process an each time 't' depends on only one probability distribution. This is called stationarity. Luckily, the techniques we will be focusing on require only Wide Sense Stationarity: the first and second moments must be time invariant.

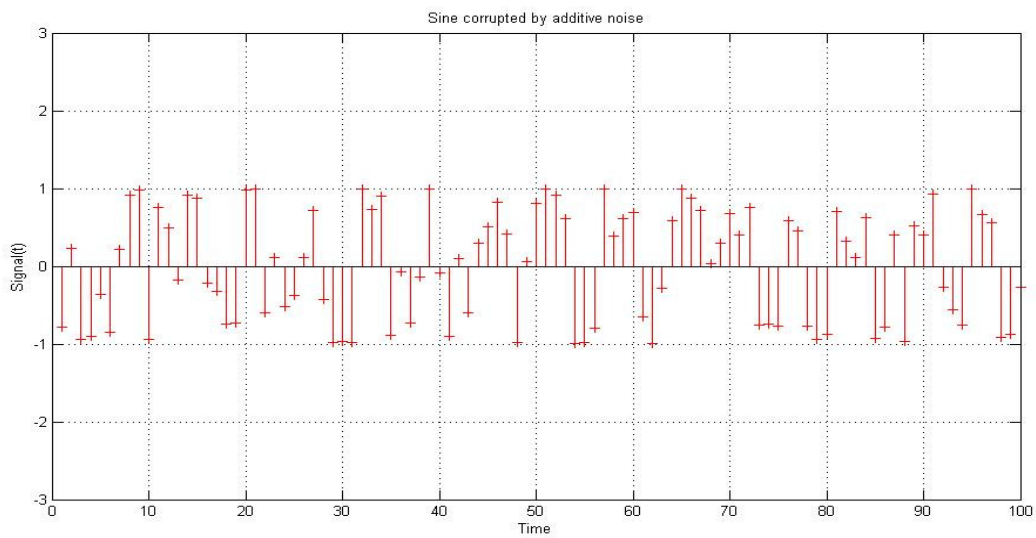
$$E(X_t) = E(X_{t-j})\ and\ E({X_t}^2) = E({X^2}_{t-j}), for\ all\ t\ and\ j = 1, \dots, t$$

Autocorrelation will be a familiar term to students of time series. A single realization of a stochastic process can be considered time series data. In which case the autocorrelation of a value of the stochastic process at time t with its jth lag is defined as:

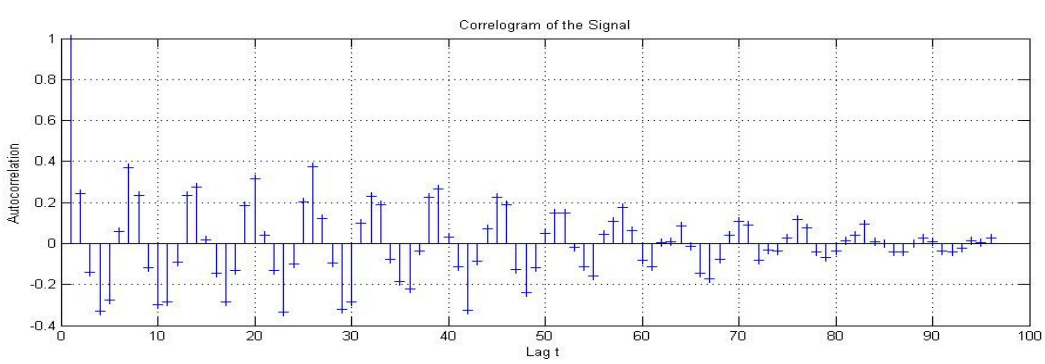
$$Corr(X_t, X_{t-j}) = \frac{Cov(X_t, X_{t-j})}{\sqrt{Var(X_t) \times Var(X_{t-j})}}$$

Autocorelation is a ubiquitously used tool in timeseries analysis because it allows us to visualize periodicities. In particular, a graph of the lag vs. autocorrelation graph, popularly known as the correlgram illustrates periodicity in the data. To better portray this idea, let us take an example of a signal: a sine wave whose angular frequency is has been corrupted by additive noise:

$$signal\colon\ X(t_i) = \sin(t_i + \mathbf{n}_{t_i}), \text{where } \mathbf{n}_{t_i} \text{ Gaussian noise}$$



The signal seems to be randomly alternating between positive and negative values. Obviously, since we have stated the signal in functional form, we know that it is not completely random and contains strong periodicity in it. Suppose we were unaware. Then, one measure we would use is the autocorrelation graph:



Notice that the autocorrelation with increasing lag order reveals a diminishing sinusoidal relationship between the present value of the function and its' jth lag. This is a tell tale sign of periodicity in the data. Unfortunately, we can go no further in our analysis of periodicity in the signal; we have reached the limits of analysis in the time domain. So how do we proceed? In order to answer that question, we must combine what we have learned about the Fourier transform with what we have learned about stochastic processes.

The Spectrum: Power Spectral Density and basics of Power Spectral Estimation

So far, we have developed the Fourier transform and important properties of stochastic processes, particularly the autocorrelation. Now, we will fuse these two concepts together to develop an extremely well respected method of analyzing stochastic processes, known as Spectral Analysis.

Spectral Analysis primarily rests on what is called the *power spectrum*. Papoulis defines the power spectrum—also known as the spectral density—of a WSS (Weak Sense Stationary) process $X(t)$ as the Fourier transform $S(w)$ of its autocorrelation, $R(t)$:

$$S(w) = \int_{-\infty}^{\infty} R(t)e^{-iwt}$$

And the autocorrelation of a function is defined, via the inversion formula of the Fourier transform, as:

$$R(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(w)e^{iwt}, (Papoulis, 412 - 416)$$

The above are known as the Wiener –Khintchine relations, after Norbert Wiener and Aleksandr Khinchin (Simon, 71).

What is it that the power spectrum illustrates? Technically speaking, it contains the exact same information as the autocorrelation function, but it is packaged differently. Recall that the Fourier transform of a time-valued process returns the process in frequency space. This is exactly what the transform does to the autocorrelation function. The power spectrum describes to what degree each frequency is present in the data, ie, at what frequencies the data are most correlated. As Meko describes, "the spectrum of a time series is the distribution of [autoco]variance of the series as a function of frequency" (Meko, chapter , pg 1).

If the autocorrelation and the spectrum contain the same information, one might ask why spectral analysis is so potent of a method. The answer is that there are many phenomena—biological,physical, and economics—which have variability that is frequency dependent, so understanding the frequency dependence is paramount. For instance, the variable costs of a business may be dependent on the seaonsality of sales or business operations that occur every so often a year. Another example is the relationship between unemployment and inflation. These are but two example of many. Spectral analysis allows us to detect and describe cyclic patterns in a stochastic process.

Spectral Estimation:

The final segment on spectral analysis deals with the problem of estimation. If we have reason to believe that there are cyclic properties in a stochastic process, can we estimate what the power spectrum will be from a single realization of the process,ie a finite number of data points? The answer is yes; in fact, there are a gamut of techniques available for doing this. We are going to briefly summarize a classical, nonparametric technique known as the periodogram.

Papoulis defines the periodogram of a signal as the process:

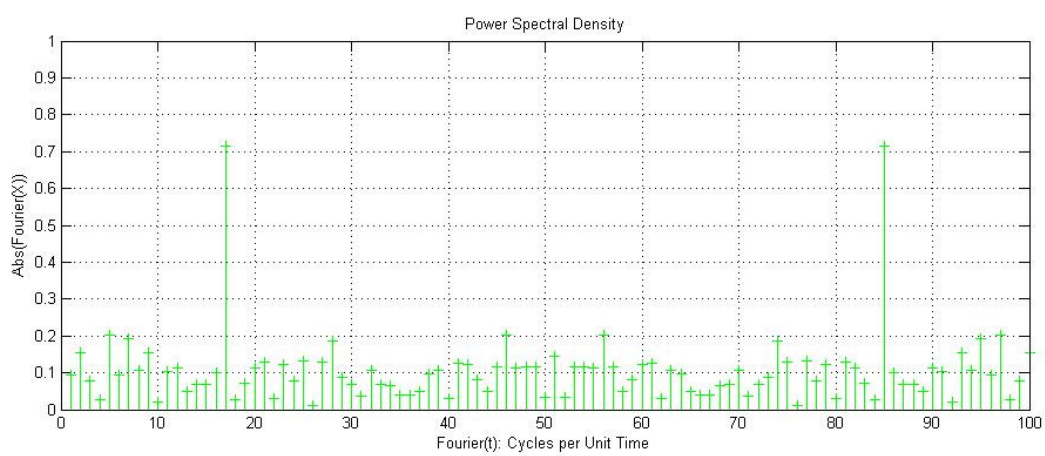
$$P(w) = \frac{1}{2T}\left|\int_{-T}^{T} X(t)e^{-iwt}dt\right|^2, (Papoulis, 539)$$

This has a discrete form as well, which is necessary if computers are to be used in evaluating the periodogram. Vaseghi defines the dsicrete form as:

$$\frac{1}{T}\left|\sum_{k=0}^{T-1} X(k)e^{-iwk}dt\right|^2, (Vaseghi, 279)$$

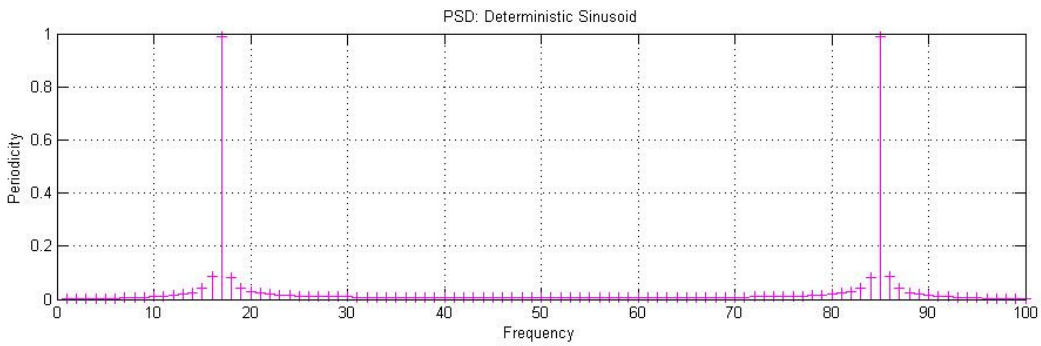
Storey notes that one can define the power spectrum as the absolute value of the Fourier coefficients, which will provide us with the total amount of information contained at a given frequency: the square of the absolute value is considered the power of the signal (Storey, 10). Because of its algorithmic simplicity, an estimation of the PSD using the periodogram is used in the MATLAB files.

At this point, we are going to continue our simple example of a sinusoid corrupted by additive noise. Taking the autocorrelation data from above and applying the Fourier transform to it, we produce the following power spectrum of the process:



The sinusoidal shape of the correlogram is completely gone. The reason for this is that frequency is now the horizontal axis. Each stem represents the magnitude of periodicity associated with each frequency. Notice that the magnitude of periodicity is extremely high at values close to 17 and 85 (this represents the true periodic nature of the sine wave). There are also small levels of periodicity that occur at every

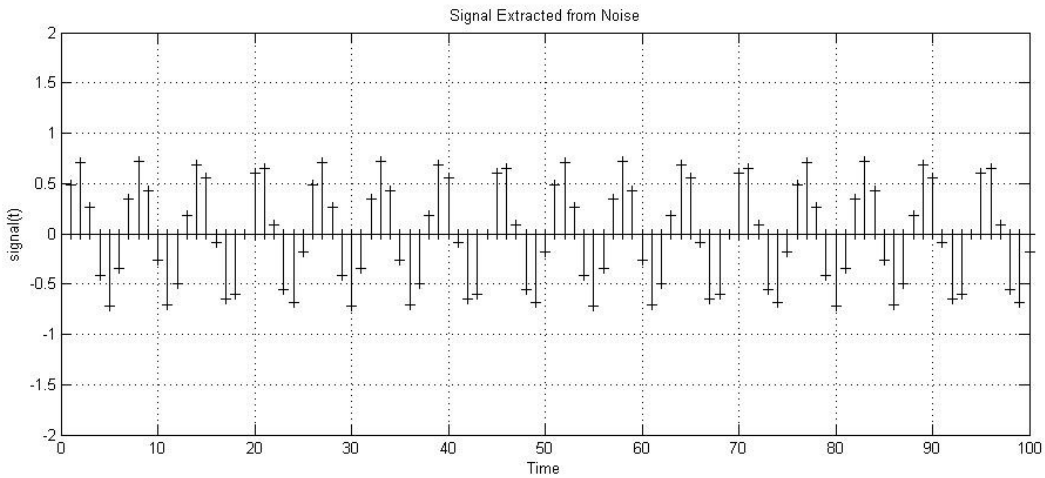
other frequency. In spectral analysis of stochastic processes, this always happens. These "fake periodicities" arise from the random variations that occur in the process. One can distinguish between the fake periodicities and the true frequencies of periodicity by noting that the additive noise occurs throughout the process, and so it should indicate a relatively even and low level of periodicity across all frequencies. The magnitude of the real periodicity, which occurs only at a few frequencies, is much larger. To illustrate this point, compare the above PSD of the random signal with that of a deterministic sinusoid:



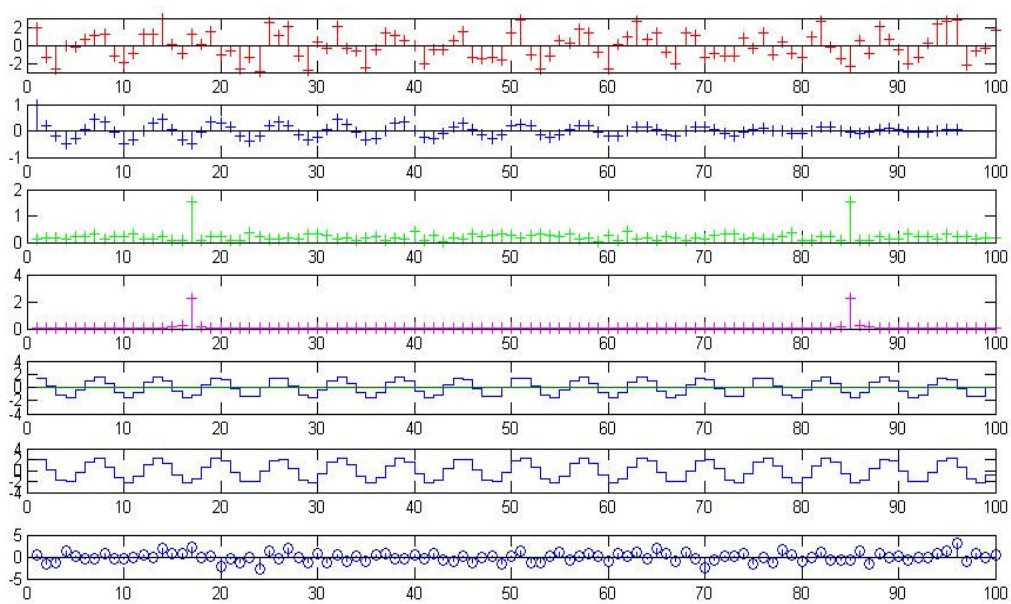
The estimated PSD of the signal and the sinusoid reflect a large amount of correlation occurring at the same frequency, but the sinusoid does not have the minute levels of periodicity at all other frequencies. This is because those periodicities are caused by the noise. This fact allows us to do something in the spectrum which is impossible to do in the correlogram: filter out the noise. If we can detect the noise by its low-level of periodicity across all frequencies, we can cut out all the data points that correspond to a spectral value that is below a benchmark.

The algorithm can be described fairly simply in words as well. In essence, we want to define a new signal. This new signal's power spectrum is equal to the old signal's power spectrum only at the frequencies at which the old spectrum's value was above 0.5. If you look at the PSD of the old signal, you'll notice that only the frequencies at which we believe the true periodicity is occurring are above this value.

The next step is to rely on the Fourier inversion formula developed in the first part of this paper to convert this power spectrum into the signal. When we do this, we obtain the following:



This has periodicities occurring at the exact same time as an uncorrupted sine function. The above example was really simple, but this technique can be applied to much more complicated signals that are corrupted by more complicated noise. To show this, we provide the entire process applied to a linear combination of trigonometric functions with their own additive noise:



From top to bottom: A signal corrupted by additive noise; the autocorrelation of the signal; the estimated Power Spectrum of the signal; the signal retrieved from noise; a graph of an uncorrupted version of the original periodic function; the noise that was filtered out. Notice that the periodicities of the uncorrupted signal and the extracted signal occur at the exact same times.

This ends our brief introduction to spectral estimation techniques of stochastic processes. If a mathematician is tasked with analyzing a stochastic process—be it variable costs, inflation, position of pollen particles, etc.—a good place to start is with a spectral analysis, because if periodicities do exist in

the process then spectral techniques will detect it. Of course, the breadth of such techniques goes far beyond the basics that were discussed above; nonetheless they serve the purpose of illustrating the power of spectral estimation.

Part II: Bayesian Estimation

Introduction to Bayesian estimation:

The spectral estimation techniques covered above are standard techniques to process a random signal. However, modern signal processing is gradually being reconceptualized in the Bayesian framework (Candy 2009). Bayesian methods are capable of providing estimations of a stochastic process that is not particularly periodic. Since Rev. Bayes' papers were posthumously published, a fissure has existed in the discipline of statistical inference. Oddly enough, the contentious issue is philosophical, and makes for a good entry point for the mathematics that will be covered in this section.

The issue is this: in mankind's scientific endeavor to understand his universe, can he ever discover its various parameters, or are the parameters themselves just abstractions? Classical probabilists (known as frequentists) believe that the mean of a stochastic process, for instance, is real-but unknown and unknowable—and can only be estimated from the data. Bayesians, on the other hand, believe that only the data is real the population mean is simply a man made abstraction; some values of this abstraction are more believable than others, given the data.

How does this relate to signal processing? Let us look at Bayes' theorem, which serves as the foundation of all Bayesian analysis. Given a set of random variables X, noisy data Y, Bayes' theorem is:

$$\Pr(X|Y) = \frac{\Pr(Y|X) \times \Pr(X)}{\Pr(Y)}$$

Where the Pr(X|Y) refers to the conditional probability "Probability of X given Y" and similar for Pr(Y|X). The probability Pr(X|Y) is called the *posterior* distribution (probability of the random variable X = x after new data has been collected). The probability Pr(Y|X) is called the likelihood; those familiar with mathematical statistics will understand that the likelihood function is a function of the parameters X for fixed observations Y. The probability Pr(X) is the prior distribution, ie, the density function that the Bayesian assigns to the random vector X before there is data. Finally Pr(Y) is known as the "evidence" but is really there to normalize the posterior density (ensuring that the integral over the complete space

is 1 (Candy, 2009). To make sense of why Bayes' theorem is important to the processing of random signals, it is helpful to first consider it as

$$\Pr(X|Y) = \frac{\Pr(Y|X)}{\Pr(Y)} \times \Pr(X)$$

In other words, Bayes' theorem adjusts our current understanding of the density of X by multiplying it by the likelihood function divided by the evidence. If it is likely that the set of observations Y would be observed if X equals a given value in its range, but unlikely that the observations Y would be observed under the Prior distribution, then the multiplying factor is large and increases the prior. The converse is also true. Thus, Bayes' theorem allows statisticians to estimate an unknown distribution as new evidence presents itself.

Sequential Bayesian Estimation

The above form of Bayes' rule applies to a general set of variables X and observations Y. However, we are interested in sequentially (in-time) estimating the posterior distribution. How do we apply this to the analysis of a stochastic process? Recall from our definition of a stochastic process that it is a sequence of random variables X_t which differ as time, t, changes. So, let us call the hidden parameters of our stochastic process a random signal that is actually "unobserved," X_t and our set of observed values of this process Y_t. We are interested in estimating:

$$\Pr(X_t|Y_t) = \frac{\Pr(Y_t|X_t)}{\Pr(Y_t)} \times \Pr(X_t)$$

From probability theory, we know that by the Law of Total Probability,

$$\Pr(Y_{t)} = \int \Pr(Y_t|X_t) \Pr(X_t) dX_t$$

Which is integrated over the entire parameter space of X_t. This turns Bayes' Theorem into the following:

$$\Pr(X_t|Y_t) = \frac{\Pr(Y_t|X_t)}{\int \Pr(Y_t|X_t) \Pr(X_t) dX_t} \times \Pr(X_t)$$

As Candy notes, this is conceptually appealing but leaves us with few ways to analytically evaluate the posterior: only a small class of priors and likelihood distributions exist such that the calculation of the posterior distribution is analytically tractable (Candy, 30-50). Candy also asserts that the large dimensionality of the integrals involved cause most numerical integration techniques to break down as well. We will deal with this daunting issue in the later discussion on Monte Carlo Markov chains, importance sampling, and particle filtering. The crucial development is that we can formulate the density of the state of an anonymous stochastic process. Now we must further develop Bayes' theorem into a method of processing random signals.

Joint Posterior Estimation:

The posterior above has the entire n-dimensional vectors of the random variable and the data. This means that values of X and Y for all time values are in this vector. After making a few assumptions, we will be able to formulate the posterior in terms of values of the random signal for a specific t*th* time. It is imperative that a step by step development of sequential Bayesian estimation model is offered so as to provide an exact understanding of how mathematicians can begin to approximate the density of a stochastic process at a given interval of time. Note that this development of the Bayesian estimation model is largely adopted from Candy, 2009

The first development is the use of the Probability chain rule (see appendix). We will also make the fundamental assumption in Bayesian signal processing that random signal X is a Markov process. Finally, we assume that the data Y are conditionally independent of the past dynamic variables.

Let's begin by reformulating the likelihood function $\Pr(Y_t|X_t)$ by taking note of the above assumptions. We can extract the tth term from the likelihood function:

$$\Pr(Y_t|X_t) = \Pr\big(y(t)\big), Y_{t-1}|x(t), X_{t-1})$$

Commas are used to denote intersections of events. From here, via the probability chain rule, this becomes:

$$\Pr(Y_t|X_t) = \Pr\big(y(t)\big), Y_{t-1}|x(t), X_{t-1}) \Rightarrow \mathbf{Pr}(\boldsymbol{y(t)}|\boldsymbol{Y_{t-1}}, \boldsymbol{x(t)}, \boldsymbol{X_{t-1}}) \times \mathbf{Pr}(\boldsymbol{Y_{t-1}}|\, \boldsymbol{x(t)}, \boldsymbol{X_{t-1}})$$

But remember we assume that y(t) is conditionally independent of the past values, namely Y_{t-1} and X_{t-1}. Then the first and second term of the above equation become

$$\Pr(y(t)|Y_{t-1}, x(t), X_{t-1}) \longrightarrow \mathbf{Pr}\big(\boldsymbol{y(t)}\big|\boldsymbol{x(t)}\big) \text{ and } \Pr(Y_{t-1}|\, x(t), X_{t-1}) \longrightarrow \mathbf{Pr}(\boldsymbol{Y_{t-1}}|\, \boldsymbol{x(t)},)$$

Making the final expression of the likelihood function:

$$\Pr(Y_t|X_t) = \Pr\big(y(t)\big|x(t)\big) \times \Pr(Y_{t-1}|\, X_{t-1})$$

We follow the exact same procedure to re-formulate the Prior distribution Pr(X_t). First extract the tth term and then employ the probability chain rule:

$$\Pr(X_t) = \Pr\,(x(t), X_{t-1} = \Pr(x(t)|X_{t-1}) \times \Pr\,(X_{t-1})$$

But recall that we assume that the random signal X is a Markov chain, so the tth term would only depend on the (t-1) term:

$$\Pr(x(t)|x(t-1)) \times \Pr\,(X_{t-1})$$

The evidence term is derived similarly: we'll skip this derivation as the reader likely understands the method now:

$$\Pr(Y_t) = \Pr(y(t), Y_{t-1}) = \Pr(y(t)|Y_{t-1}) \times \Pr\,(Y_{t-1})$$

Now that we have successfully reformulated our three distributions using Bayes', the chain rule, and assumptions about independence, we can plug them back together into Bayes' definition of the posterior, which yields the following messy definition of the posYtt-1erior distribution:

$$\Pr(X_t|Y_t) = \frac{[\Pr\,(Y_{t-1}|X_{t-1}) \times \Pr\,(y(t)|x(t))]\left[\Pr\big(x(t)\big|x(t-1)\big) \times \Pr\,(X_{t-1})\right]}{\Pr(y(t)|Y_{t-1}) \times \Pr\,(Y_{t-1})}$$

Seemingly, we have made a relatively simply definition of our desired density even more complicated. But notice that Bayes' theorem defines the posterior distribution at the previous time period, t-1, by:

$$\Pr(X_{t-1}|Y_{t-1}) = \frac{\Pr(Y_{t-1}|X_{t-1})}{\Pr\,(Y_{t-1})} \times \Pr\,(X_{t-1})$$

This definition is embedded in the above messy definition of the posterior at the current time. We now extract those terms and represent them by $\Pr(X_{t-1}|Y_{t-1})$. This yields:

$$\Pr(X_t|Y_t) = \left[\frac{\Pr\big(y(t)\big|x(t)\big) \times \Pr\,(x(t)|x(t-1))}{\Pr\,(y(t)|Y_{t-1})}\right] \times \Pr\,(X_{t-1}|Y_{t-1})$$

This definition explains that if we have the value of the posterior at the previous time, we can evaluate it at the current time by multiplying it by a weighting function, which is in brackets. We can think of our initial prior distribution as our first value of $\Pr\,(X_{t-1}|Y_{t-1})$. Then, as t increases, we gain more recent data that will change our evaluation of the joint posterior; for each value of t, the posterior is adjusted by the weighting function to yield a more current evaluation of it. In other words:

$$New = Weight \times Old$$

$$\Pr(X_t|Y_t) = W(t, t-1) \times \Pr\,(X_{t-1}|Y_{t-1})$$

Note that all the terms in the weighting function come from either the likelihood, the prior, or the evidence distribution which we have already stated are easy to attain. In fact, the new formulation is still a form of Bayes' theorem, just manipulated for more convenient use.

The last part of the development of a useful, sequential Bayesian estimator results from the following observation. While we have successfully created a form for the *joint posterior*, this is actually a joint density of each x(t) for every t. Knowing such a density is not as useful to us as the density of only the value of the stochastic process at the next time interval. In other words, we must find the marginal density from the above joint density. This will allow us to predict the value of the stochastic process.

Precisely, our goal is to estimate Pr(x(t)| Y_t), the current value of the stochastic process given all available data at time t. Let us start by specifying our prediction of the stochastic process in the next time interval. Specifically,

$$\Pr(x(t)|Y_{t-1}) = \int \Pr(x(t), x(t-1)|Y_{t-1})\, dx(t-1)$$

Applying Bayes' rule to the integrand and then realizing the Markovcian assumption of conditional independence between X and Y, the following is achieved:

$$\begin{aligned}\Pr(x(t), x(t-1)|Y_{t-1}) &= \Pr(x(t)|x(t-1), Y_{t-1}) \times \Pr(x(t-1)|Y_{t-1}) \\ &= \Pr\,(x(t)|x(t-1)) \times \Pr\,(x(t-1)|Y_{t-1})\end{aligned}$$

So our prediction of the value of the process at the next time interval, given the data up to the present, is:

$$\Pr(x(t)|Y_{t-1}) = \int \Pr\,(x(t)|x(t-1)) \times \Pr\,(x(t-1)|Y_{t-1}))$$

We then use the definition of conditional probability to reformulate the marginal posterior distribution:

$$\Pr(x(t)|Y_t) = \frac{\Pr\,(x(t), Y_t)}{\Pr\,(Y_t)} = \frac{\Pr\,(\mathrm{x(t)},\ \mathrm{y(t)}, \mathrm{Y_{t-1}})}{\Pr\,(\mathrm{y(t)}, \mathrm{Y_{t-1}})}$$

Where we just applied the definition of conditional probability twice. Now, applying Bayes' theorem to the right-most expression, we get:

$$\Pr(x(t)|Y_t) = \frac{\Pr\,(y(t)|x(t), Y_{t-1}) \times \Pr\,(x(t)|Y_{t-1}) \times \Pr\,(Y_{t-1})}{\Pr\,(\mathrm{y(t)}|Y_{t-1}) \times \Pr\,(Y_{t-1})}$$

Cancelling like terms, realizing again the Markovian assumption, the following is achieved:

$$\Pr(x(t)|Y_t) = \frac{\Pr\,(y(t)|x(t)) \times \Pr\,(x(t)|Y_{t-1})}{\Pr\,(\mathrm{y(t)}|Y_{t-1})}$$

Thus, the posterior distribution of the next value of the stochastic process is an *update* of the previous posterior $\Pr\,(x(t)|Y_{t-1})$, which was used to predict the value at the current time. Like in the joint case, we can consider the marginal posterior, or *update* distribution, as an adjustment of our prediction of the marginal adjusted by some weighting function which takes into account the new data:

Update = Weight X Prediction

$$\Pr(x(t)|Y_t) = W_c(t, t-1) \times \Pr(x(t)|Y_{t-1})$$

Where weight is defined as:

$$W_c(t, t-1) = \frac{\Pr\,(y(t)|x(t))}{\Pr\,(y(t)|Y_{t-1})}$$

This completes our development of an on-line, sequential Bayesian estimator of an anonymous stochastic process. The model is recursive, in that future values are dependent on previous values and a weighting function only. Starting with some prior distribution, the sequential Bayesian estimator self-corrects its density of the process's next value as new data comes in. This entire model works within the Bayesian framework, in that a previously un-estimatable distribution are broken down into a likelihood,

a prior, and an evidence distribution, all of which can be attained. This recursive Bayesian estimation model is a framework that can be expanded into many familiar state-space model techniques, such as the kalman filter and its' nonlinear extensions (unscented and extended kalman filter, etc.). This is one reason why understanding the general Bayesian estimation model is so important (for more information on the Kalman filter, see Simon or Candy).

However, one significant problem lingers. Even though Bayes' theorem gives us a way of describing the posterior distribution, very rarely does a closed form solution exist (remember, Bayes' theorem asks us to compute integrals of density's which can become very complicated in real world scenarios). The next section will examine the methods available to deal with problem, methods which can be generally classed as Monte Carlo simulation techniques.

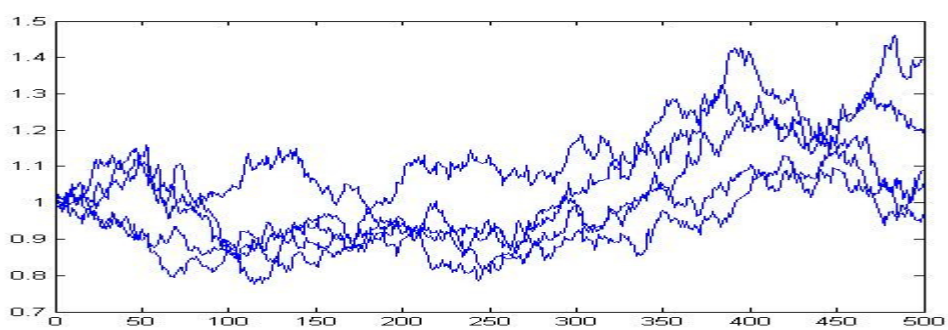
Monte Carlo Simulation:

First, we will offer a general explanation of what Monte Carlo simulation is. The idea is that if we cannot evaluate a density—or any integral-- analytically, we can create a computer program that will simulate the function at values in its domain that are chosen randomly. If we run this simulation 1000s of times, we will get a histogram of possible values of the function which approximates the density.

Here is an interesting example from quantitative finance. The return on a stock at the end of a trading day is thought to follow a certain stochastic differential equation whose solution is a stochastic process, given by:

$$S(T) = S(0)e^{\left(\mu - \frac{1}{2}\sigma^2\right)T + \sigma B(T)}$$

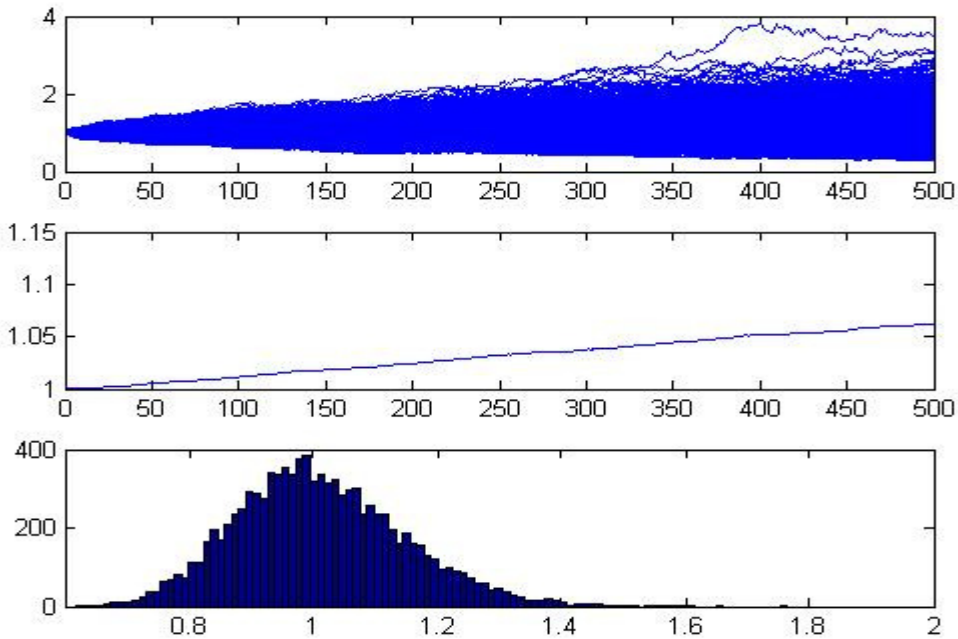
In this case the integral can be evaluated analytically but, as stated above, the solution to a stochastic differential equation is still a stochastic process. This means that the above solution will generate a different graph each time it is computed. To illustrate this, the following Matlab graph is provided. Note that coefficients were arbitrarily chosen.



5 different paths of the solution

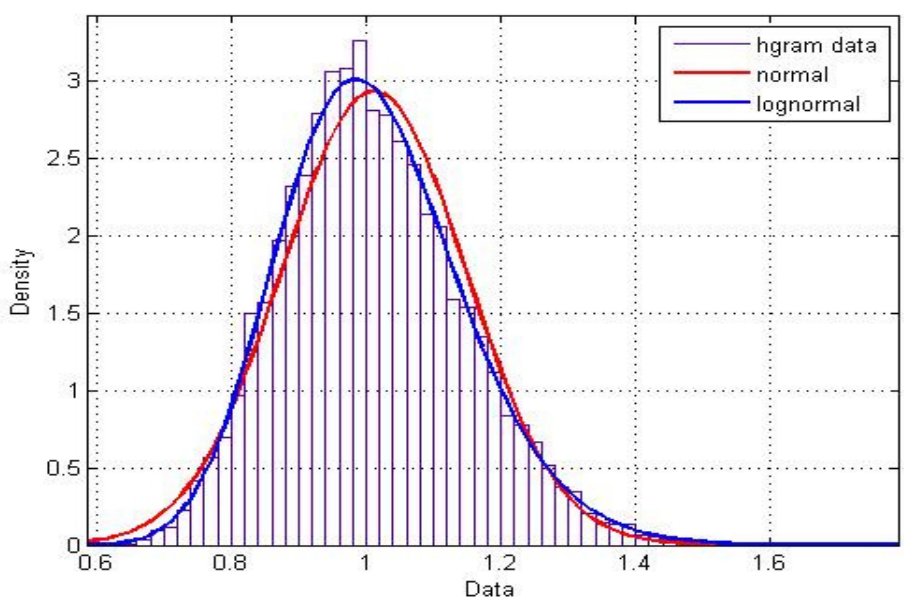
The graph shows 5 different plausible paths of the same solution. To navigate this obstacle, mathematical modelers traditionally employ Monte Carlo techniques. In essence, the solution is simulated thousands of times and then averaged. Matlab allows us to do just this. The following Matlab

graphs show ten thousand sample paths, the mean sample path, and a histogram of S(T) at the end of its trajectory—the point which represents price at the end of a trading day.



The simulated paths, the mean path, and a histogram of the function at the end of the time interval

Because each simulation of the solution is independent and identically distributed, the central limit theorem suggests that the distribution of multiple simulations will be normal if the number of simulations is sufficiently large. This is showed by fitting a density curve on the histogram:



Fitting a density curve to the histogram from the last figure. The central limit theorem is verified, as normal distributions are the best fit.

The best fits are normal and lognormal. Note that lognormal is also accurate because the solution is in the form of the constant **e** raised to a power which contains the random variable. Using Matlab to fit this density onto the data allows us to extrapolate a number of statistical properties of the stock price. For instance, it gives us an approximate mean and standard deviation. With this information, we can begin to formulate a plausible trading algorithm, one that buys stocks based on the likelihood of a given stock's ending price being higher than the price at which it was bought.

In this example, a stochastic process was simulated thousands of times to achieve an approximate density of its value. In Bayesian estimation, the goal will be slightly different, since it is the density itself which cannot be solved analytically. Nonetheless, the basic strategy is the same: we will generate random points from the domain of the density, propagate them through the density function (in other words, evaluate the density at those points), and create a histogram of this process after it has been repeated thousands of times.

Mathematical Development of Monte Carlo Methods

Candy crystallizes our objective with Monte Carlo when he states, "The goal of Bayesian techniques using MC methods is to generate a set of independent samples from the target posterior distribution with enough samples to perform accurate inferences" (Candy 2009). We are used to the idea of generating random samples from a given distribution, but the reverse is also true: we can approximate the distribution via samples from another distribution (in this case, from the prior distribution). Consider some density f(x). Then we know that the expectation of a function g(x) is:

$$E\big(g(x)\big) = \int g(x) \times f(x)dx$$

Now suppose we passed an *n* amount of values from g(x)'s range—which we will call samples—through g(x), then the sample mean would be the Monte Carlo estimate of E(g(x)):

$$\breve{g}(x) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

Provided the true expectation exists, the sample mean (ie Monte Carlo estimate) and the true mean converge in probability according to the weak law of large numbers (C. Anderson 2, Alpaslan). As Anderson notes, this revelation "comes to life and becomes useful when one realizes that very many quantities of interest may be cast as expectations" (C. Anderson 2). Specifically he is speaking about probabilities, distributions, and integrals in general. We are interested in the former two. For instance, the probability that some random variable X takes on some value in a set A can be expressed as:

$$P(X \in A) = E\big(I_A(X)\big) \approx \frac{1}{n}\sum_{i=1}^{n} I_A(x_i)$$

Where $I_A(X)$ is the indicator function that takes the value 1 when $X \in A$ and 0 when it does not. If we wanted to approximate a distribution, then we would divide the entire range of the random variable into N amount of discrete intervals, and perform the above probability estimation for each interval; the results would be a sequence of values that, when represented as individual bars in a graph that are adjacent to each other, form a histogram that approximates the distribution.

Note that the evidence distribution really functions as a normalizing factor, to ensure the posterior equals 1 when evaluated across its range. Then,

$$\Pr(X|Y) \propto \Pr(Y|X) \times \Pr(X)$$

Obviously the prior is assumed to be known, so the method is simply to:

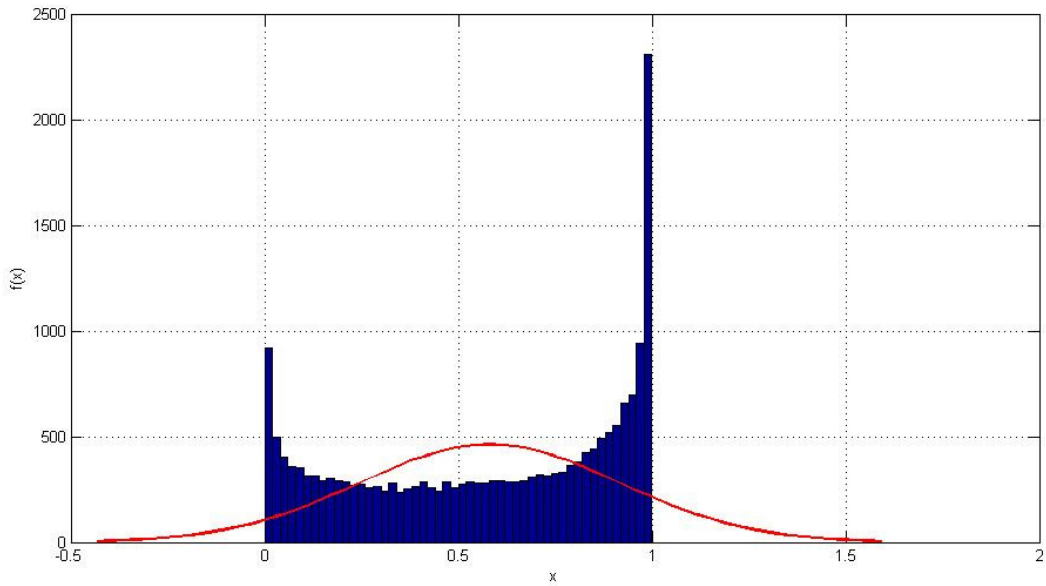
1. Generate samples from the prior
2. Pass these samples through the likelihood
3. Estimate the posterior by multiplying likelihood by prior.

This methodology is not as simple as one might think because generating good samples is a science in and of itself. Candy poses to problems related to sampling. First, how do we in fact generate samples from distributions? Second, how can we be sure that these samples are independent and identically distributed samples of the posterior?

The answer to the first question is that, with computer software such as MATLAB, we can generate pseudo-random numbers from a variety of different known distribution. But the answer to the second question is much more complicated. Consider using a uniform distribution to try and simulate the density:

$$f_X(x) = e^{-(x)^2}$$

We sample 2000 values from a uniform distribution, pass each one through the above density, and create a histogram of the values obtained:



Readers should be aware that the density being estimated should be a bell curve. So, can we just generate random numbers from a uniform distribution and propagate them through the density and expect a good approximation for large enough n? The simulation above says no. This is because there is no reason to believe these samples passed through the target density are i.i.d from the target density. The quest of developing methods that generate iid samples from a target posterior is the topic of the next section. We go over these methods, starting with rejection sampling, importance sampling, and then sequential importance sampling.

Sampling:

Rejection Sampling

Rejection sampling is a versatile sampling method that allows us to estimate a density up to a proportionality constant, which remains unknown. Rejection sampling is also important because it is related to more advanced sampling techniques that will be discussed later.

Suppose that $P_a(X)$ is the density that we are interested in, but we cannot sample from it directly because it has no analytical form. We choose another another density $P_b(X)$ which exists such that

$$P_a(X) \leq M \times P_b(X)$$

Then note that with simple algebraic manipulation, the follow is made obvious:

$$\frac{P_a(X)}{P_b(X)} \leq M \text{ and } \frac{P_a(X)}{M \times P_b(X)} \leq 1$$

Noting this, the rejection sampling technique is as follows:

1. Generate a sample from $P_b(X)$, denoted as x_k
2. Then generate a sample from a uniform distribution, $u_k \in U(0,1)$. Naturally, $0 < u_k < 1$
3. If $u_k \leq \frac{P_a(X)}{M \times P_b(X)}$, then accept the sample as a sample of $P_a(X)$: inotherwords the ith sample of $P_a(X)$ is equal to x_k if $u_k \leq \frac{P_a(x_k)}{M \times P_b(x_k)}$.

This sampling method is more effective because it makes use of density related to our target. Determining a related density, however is a nontrivial endeavor (Candy 2009, 63). Because our interest in rejection sampling is its use in more advanced method, we will not explore this problem.

Importance Sampling:

A related sampling method is called Importance Sampling and its on-line relative, Sequential Importance Sampling. It is simply another way to mitigate difficulties with the inability to directly sample from a posterior distribution. To understand it, let us return to our discussion about expectations.

Essentially everything we are probabilitistically interested in can be thought of as an expectation. Say that we want to estimate the posterior density f(x) and we know g'(x) where g(x) = C*g'(x); in other words, we know g(x) up to a constant. Then a probability of x can be written as:

$$\Pr(X) = \int f(x)dx$$

As Anderson writes, importance sampling rests on multiplying this probability by 1, which is dressed up in a tricky fashion: 1 = $\frac{g'(x)}{g'(x)}$. Then

$$\Pr(X) = \int f'(x)dx = \int f'(x) \times \frac{g'(x)}{g'(x)} dx$$

But, recalling the definition of expectation, we realize that:

$$\int f'(x) \times \frac{g'(x)}{g'(x)} dx = E_{g'(x)}\left[\frac{f'(x)}{g'(x)}\right]$$

This expectation can be estimated using the Monte Carlo estimator, where samples are generated from the known g'(x) distribution:

$$E_{g'(x)} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{f'(x_i)}{g'(x_i)} \cong \Pr(X)$$

We can estimate the entire distribution just as we did before, by dividing the range of the random variable into discrete intervals and applying the Monte Carlo estimate to each interval (Anderson). Of course, we usually need the density to calculate some value such as the mean or variance, which are functions of the density. If such is the case, we can simply modify the procedure like so:

Suppose it is the posterior distribution, $\Pr(X_t|Y_t)$, that we are trying to sample for in order to find the expectation of a function of x, $f(x)$. The importance distribution is $g'(x)$. Then:

$$\int f(x) \times \Pr(X_t|Y_t) \times \frac{g'(x)}{g'(x)} dx = E_{g'(x)}\left[\frac{f(x) \times \Pr(X_t|Y_t)}{g'(x)}\right]$$

Now say we apply Bayes' theorem to the posterior and define a weighting function such that:

$$\int f(x) \times \breve{W}(t) \times g'(x) dx$$

Where $\breve{W}(t) = \frac{\Pr(X_t|Y_t)}{g'(x)} = \frac{\Pr(Y_t|X_t) \times \Pr(X_t)}{\Pr(Y_t) \times g'(x)}$. But in practice, as has been discussed above, we rarely know the evidence $\Pr(Y_t)$. Candy explains a fix for this. Since the $\frac{1}{\Pr(Y_t).}$ is the issue, let us define a new weighting function that exists without it:

$$W(t) = \frac{\Pr(Y_t|X_t) \times \Pr(X_t)}{\mathrm{g}'(x)}$$

Then

$$\int f(x) \times \breve{W}(t) \times g'(x) dx = \frac{1}{\Pr(Y_t).} \int f(x) \times W(t) \times g'(x) dx$$

Where we sampling substituted the new weight and extracted the troublesome evidence. Now note the definition of the evidence probability and a simple algebraic manipulation of the definition of the weighting function:

$$\Pr(Y_{t)} = \int \Pr(Y_t|X_t) \Pr(X_t)\, dX_t, and$$

$$W(t) \times \mathrm{g}'(x) = \Pr(Y_t|X_t) \times \Pr(X_t)$$

Then we can define the normalizing constant as the integral of the weight times the g'(x). Then:

$$E_{\Pr(X_t|Y_t)}\big(f(x)\big) = \frac{1}{\Pr(Y_t).} \int f(x) \times W(t) \times g'(x) dx = \frac{E_{g'(x)}[W(t)f(x)]}{\int W(t) \times \mathrm{g}'(x)} = \frac{E_{g'(x)}[W(t)f(x)]}{E_{g'(x)}[W(t)]}$$

$$\Rightarrow \frac{E_{g'(x)}[W(t)f(x)]}{E_{g'(x)}[W(t)]} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} W_i(t) \times f(x_i)}{\frac{1}{n}\sum_{i=1}^{n} W_i(t)} = \sum_{i}^{n} \overline{W_i}(t) \times f(x_i)$$

Where the weight in the right most equation is

$$\overline{W_i}(t) = \frac{W_i(t) \times f(x_i)}{\sum_{i=1}^{n} W_i(t)}$$

But, if the expectation is defined as $\sum_i^n \overline{W_i}(t) \times f(x_i)$, then the posterior itself must therefore be:

$$\Pr(X_t|Y_t) = \sum_{i}^{n} \overline{W_i}(t) \times \delta(x - x_i)$$

Intuitively speaking, this is obvious because a discrete representation of an expectation of a function is the sum of the function multiplied by its density. Thus importance sampling gives us a density g(x) from which we can generate samples to propagate through the weighting function, which is really comprised of the likelihood and prior divided by g(x). Now, the only thing that remains is to show how this derivation can be furthered to develop an on-line, sequential Bayesian Importance sampling algorithm.

Sequential Importance Sampling:

We can now take everything we have developed above and construct a description of an algorithm for a sequential importance sampler. They key ideas are the above development of a batch importance sampler, the Markov assumption of the stochastic process, and the condition independence conditions of the measurements. Let us call the sampling distribution $q(X_t|Y_t) = q(X_{t-1}|Y_t) \times q(x(t)|X_{t-1}, Y_{t-1})$, which follows from Bayes' rule. Then the unnormalized weight, W(t), defined above becomes:

$$W(t) = \frac{\Pr(Y_t|X_t) \times \Pr(X_t)}{q(X_{t-1}|Y_{t-1}) \times q(x(t)|X_{t-1}, Y_{t-1})}$$

Now note that if we multiply the numerator and denominator by $\Pr(Y_{t-1}|X_{t-1}) \times \Pr(X_{t-1})$:

$$W(t) = \frac{\Pr(Y_t|X_t) \times \Pr(X_t)}{q(X_{t-1}|Y_{t-1}) \times q(x(t)|X_{t-1}, Y_t)} \times \frac{\Pr(Y_{t-1}|X_{t-1}) \times \Pr(X_{t-1})}{\Pr(Y_{t-1}|X_{t-1}) \times \Pr(X_{t-1})}$$

By rearranging terms, we see that:

$$W(t) = \frac{\Pr(Y_{t-1}|X_{t-1}) \times \Pr(X_{t-1})}{q(X_{t-1}|, Y_{t-1})} \times \frac{\Pr(Y_t|X_t) \times \Pr(X_t)}{\Pr(Y_{t-1}|X_{t-1}) \times \Pr(X_{t-1}) \times q(x(t)|X_{t-1}, Y_t)}$$

Which is obviously equal to:

$$W(t) = W(t-1) \times \frac{\Pr(Y_t|X_t) \times \Pr(X_t)}{\Pr(Y_{t-1}|X_{t-1}) \times \Pr(X_{t-1}) \times q(x(t)|X_{t-1}, Y_t)}$$

At this point, since we are interested in a sequential sampler, we must deal with the batch variables. Because of the Markov property of the dynamic variable, conditional independence, and the probability chain rule, we can do the following:

$$Pr(Y_t|X_t) = \Pr\big(y(t)\big|x(t)\big) \prod_{k=0}^{t-1} \Pr\big(y(k)\big|x(k)\big), and$$

$$\Pr(X_t) = \Pr(x(t)|x(t-1)) \prod_{k=0}^{t-1} \Pr(x(k)|x(k-1))$$

Where we extracted the tth term from the product. The product then, is an alternate definition of $Pr(Y_{t-1}|X_{t-1})$ and $\Pr(X_{t-1})$ Then, the equation for the weight W(t) becomes:

$$W(t) = W(t-1) \times \frac{\Pr\big(y(t)\big|x(t)\big)\prod_{k=0}^{t-1}\Pr\big(y(k)\big|x(k)\big) \times \Pr\,(x(t)|x(t-1))\prod_{k=0}^{t-1}\Pr\,(x(k)|x(k-1))}{\prod_{k=0}^{t-1}\Pr\big(y(k)\big|x(k)\big) \times \prod_{k=0}^{t-1}\Pr\,(x(k)|x(k-1)) \times q(x(t)|X_{t-1},Y_t)}$$

The products cancel, yielding:

$$W(t) = W(t-1) \times \frac{\Pr\big(y(t)\big|x(t)\big) \times \Pr\,(x(t)|x(t-1))}{q(x(t)|X_{t-1},Y_t)}$$

This defines the new weight as the likelihood of the present time multiplied by the posterior of the previous time divided by the sampling distribution. With the above equations, Candy notes that we have finally arrived at a description of an algorithm for Bayesian sequential sampling (Candy, 86):

1. Draw samples the proposal, $q(x(t)|X_{t-1},Y_t)$
2. Evaluate the likelihood and conditional distributions: $\Pr\big(y(t)\big|x(t)\big)$, $\Pr\,(x(t)|x(t-1))$
3. Calculate the unnormalized weight: W(t)
4. Normalize the weights as was done in the Batch scenario: $\overline{W}_i(t) = \frac{\mathrm{W(t)}}{\sum_{i=1}^{n} W_i(t)}$
5. Finally, estimate the posterior distribution: $\Pr(x(t)|Y_t) = \overline{W}_i(t) \times \delta(x(t) - x_i(t))$

Importance sampling requires some thought as to what proposal density to begin from, and the algorithms can be quite complicated. Nonetheless, they are extremely important in advanced Bayesian estimation techniques, one of which we will describe in the conclusion. But first, we will discuss an algorithm that has a special place in the history of Monte Carlo simulation.

Markov Chain Monte Carlo

We know that samples can be generated from the posterior by passing values from the posterior's range through posterior, for instance by using a uniform distribution. The problem is that there is no guarantee that these samples are independent, and independent samples are necessary for the laws of large numbers to work. One technique to overcome this is generate random samples to pass through the posterior via a Markov Chain. Recall from the above section on stochastic processes that a Markov chain is a discrete stochastic process whose value at the current time conditionally depends only on the process's value at the previous time interval:

$$\Pr\Big(X_i(t)\Big|X_j(t-1),\dots,X_k(0)\Big) = \Pr\,(X_i(t)|X_j(t-1))$$

A Markov chain begins with an initial distribution, $\Pr(X_i(0))$ and evolves into $\Pr(X_i(t))$ according to a "transition kernel" (Candy 2009, 70):

$$\Pr\big(X_i(t)\big) = \sum_j \Pr\Big(X_i(t)\Big|X_j(t-1)\Big) \times \Pr\,(X_j(t-1))$$

For Markov chains to be useful in Monte Carlo simulation, they must be both time reversible chains and ergodic. That means that the transition probability from $X_i(t)$ to $X_j(t-1)$ must be the same as the probability from $X_j(t-1)$ to $X_i(t)$. Markov chains with these properties can be used to sample from distributions which are analytically intractable. One of the most powerful of these methods is known as Metropolis Hastings Sampling.

Metropolis Hastings Algorithm

Monte Carlo simulation techniques were a byproduct of the United States' Nuclear Armament projects during and after World War II. Originally, these methods were used to try and predict when the bombardment of an atom by neutrinos will cause a chain reaction, leading to an explosion. Two physicists, Nicholas Metropolis and W. Keith Hastings, developed a method for obtaining a sequence of random samples from an analytically intractable probability distribution. This method, known as the Metropolis Hastings Algorithm, thus has great usefulness to Bayesian estimation.

Having discussed rejection sampling and Markov chains, developing the M-H method will be relatively simple. Essentially, we will define a Markov chain whose next value is a new sample to be passed through the target distribution (the posterior, for instance). This new sample is generated from the previous sample according to the Markov chain's transition kernel. However, before this new sample is accepted as a sample of the target distribution, it will face a acceptance/rejection test similar to in rejection sampling. First we pick a random starting point that is in the range of the target distribution:

$$x_0 \rightarrow p_x(x_0)$$

Then, we pick a proposal distribution from which to initially sample: Q($\tilde{x}_1, x_0$) (the prior?). This proposal is a markov chain in that : Q($\tilde{x}_1, x_0$) is the transition kernel of a markov chain We then generate a candidate sample from the proposal: $\tilde{x}_1 \rightarrow q(x)$. (Unclear the relationship between proposal, Markov chain). Now we must determine if this sample can in fact be considered a sample of the target. We do this by calculating a value A such that:

$$A = B \times C$$

Where

$$B = \frac{p_x(\tilde{x}_1)}{p_x(x_0)}, the\ likelihood\ ratio$$

And

$$C = \frac{Q(x_0, \tilde{x}_1)}{Q(\tilde{x}_1, x_0)}, the\ ratio\ of\ the\ proposal\ density\ in\ two\ directions$$

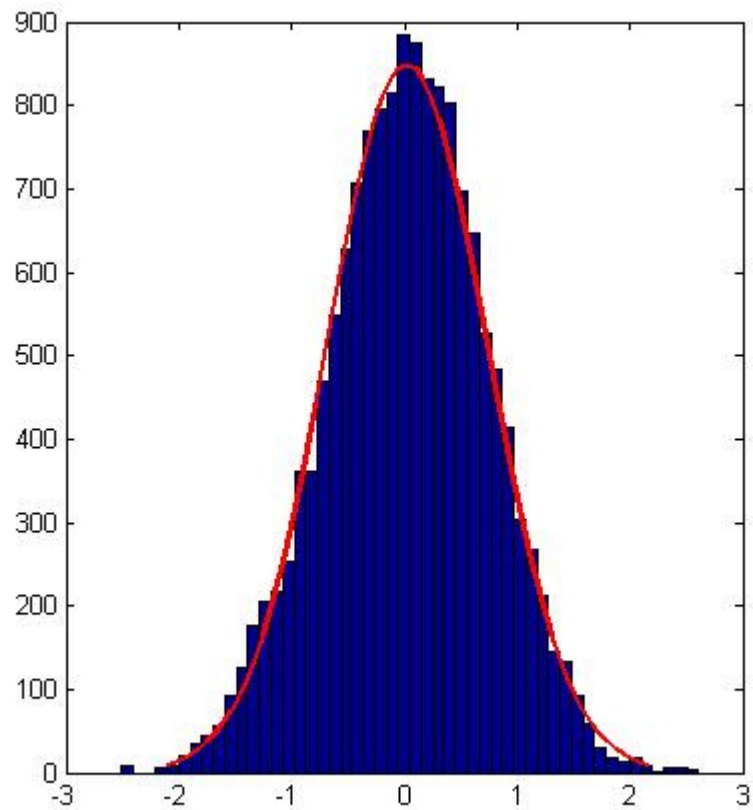
If we restricted ourselves to purely symmetric proposal densities, then C = 1 and our decision is based solely on the value of B. But in generality we speak of A. Intuitively speaking, A represents the a ratio of probabilities: therefore if it is greater than 1, the probability of the numerator is greater than the

denominator. In other words, it is more likely that $\tilde{x}_1$ is a sample of the target than x_0. If A < 1, then A is the probability that $x_1 = \tilde{x}_1$. In this case, one could either accept the candidate sample if A > 1- A, or as is generally done, generate a value *u* from a Uniform distribution U(0,1) and accept the candidate if A > u. The entire strategy can be summarized as follows: Accept the candidate sample generated from the proposal density if, given $\alpha \sim U(0,1)$,

$$\alpha < min\{A, 1\}$$

So, only samples which have a higher likelihood of being a sample of the target distribution than the previous samples are accepted. The power of this method is obvious when one realizes that Markov chains converge to an invariant distribution, which in this case means that sooner or later, the likelihood of the next candidate sample will equal the likelihood of the last sample. When this occurs, every sample there after shares the same likelihood of being a sample, and these samples thereafter can be considered samples of the target distribution. Then, Monte Carlo simulation can occur as usual.

Because of the importance of this sampling technique, a MATLAB illustration is provided. In this scenario, we make the unknown target distribution $f_x(x) = e^{-x^2}$. Obviously this is a really simple case for illustration. We will pick a **uniform proposal distribution** as Candy does in his illustration.



Once the Markov chain reached its invariant distribution, the script sampled 15,000 times more, the results of which are plotted in the above histogram. Notice that the distribution looks Gaussian, which we would expect. This verifies that the sampling method works and is superior to the ineffective, direct sampling from a uniform distribution. In practice, it would be a more complicated posterior that we are estimating: it would be comprised of multi-dimensional integrals (arising from the prior and likelihood multiplication), the answer of which is not known analytically.

Conclusion

This paper has completed its overview of the Spectral and Bayesian estimation frameworks for stochastic process analysis. The Spectral techniques are extremely well suited for detecting periodicities in a stochastic process, and periodicities give analysts a substantial tool for prediction. But what if a stochastic process is aperiodic? Then Bayesian estimation provides a robust framework for making dynamic, sequential predictions. Bayesian estimation is grounded in simulation, which requires effective sampling techniques. The development of such techniques has been quite lengthy, but hopefully has illustrated how the future value of a stochastic process can be given a probabilistic framework, ie, an analysis that provides a density function of its next value. Using the mathematics and sampling techniques explained, very complicated algorithms can be set up to estimate the value of a stochastic process dynamically, ie, as more current data comes in.

The methods discussed in this paper will hopefully provide readers with the foundational knowledge to go further in the analysis of stochastic processes. I personally come from an economics' background. Generally, an undergraduate economics education contains nothing about stochastic processes; economics education at the level is unduly concentrated on regression analysis. For many economic variables, such as inflation, unemployment, price trends, etc., regression analysis is not nearly robust or reflective of the probabilistic nature of the phenomena. For this reason, a solid understanding in the basics of stochastic process analysis is critical. With luck, this paper has provided such an understanding of the basics.

All MATLAB code was created by the author using MATLAB, the Econometrics Toolbox, and the Signal Processing Toolbox

Bibliography

Attaway, Stormy. MATLAB: A Practical Introduction to Programming and Problem Solving. Elsevier, Inc: Oxford, UK (2009).

Candy, James. Bayesian Signal Processing: Classical, Modern, and Particle Filtering Methods. Wiley Publication: Hoboken, NJ (2009).

Kay, Steven. Intuitive Probability and Random Processes using MATLAB. Springer Publication (2006).

Lay, David. Linear Algebra and its Applications, Update. Third Edition. Pearson: Boston, MA (2006).

Meko, David. Applied Time Series Analysis. http://www.ltrr.arizona.edu/~dmeko/geos585a.html#cLesson8

Papoulis, A. Probability, Random Variables, and Stochastic Processes. Fourth edition. McGraw Hill Publication: New York, NY (2002).

Simon, Dan. Optimal State Estimation; Kalman, H-infinity, and Nonlinear Approaches. Wiley Publication: Hoboken, NJ (2006).

Storey, Brian. Computing Fourier series and Power Spectrum with MATLAB. http://faculty.olin.edu/bstorey/Notes/Fourier.pdf

Vaseghi, Saeed. Advanced Digital signal Processing and Noise Reduction. Fourth edition. Wiley Publication: West Sussex, UK (2008).

Wikipedia. Fourier Transform. http://en.wikipedia.org/wiki/Fourier_Transform

WolframMathWrold. Fourier Series. http://mathworld.wolfram.com/FourierSeries.html