Overview of the Theory of Probability and Statistics

Probability is the branch of mathematics that studies the possible outcomes of given events together with the outcomes' relative likelihoods and distributions. In common usage, the word "probability" is used to mean the chance that a particular event (or set of events) will occur expressed on a linear scale from 0 (impossibility) to 1 (certainty), also expressed as a percentage between 0 and 100%. The analysis of events governed by probability is called **statistics**.

A properly normalized function that assigns a probability "density" to each possible outcome within some interval is called a **probability function** (or probability distribution function), and its cumulative value (integral for a continuous distribution or sum for a discrete distribution) is called a **distribution function** (or cumulative distribution function).

A **variate** is defined as the set of all **random variables** that obey a given probabilistic law. It is common practice to denote a variate with a capital letter (most commonly X). The set of all values that X can take is then called the range, denoted $\Re_X$. Specific elements in the range of X are called quantiles and are denoted x. The probability that a variate X assumes the element x is denoted $\mathrm{P}(X = x)$.

Distribution Function

The **distribution function** $F(x)$, also called the **cumulative distribution function** (CDF) or **probability distribution function**, describes the probability that a variate X takes on a value less then or equal to a number of x.

$$F(x) = \mathrm{P}(X \le x)$$

A distribution function satisfies

$$F(-\infty) = 0\,,\ F(+\infty) = 1$$

$$0 \le F(x) \le 1$$

There exist distributions that are neither continuous nor discrete.

Given continuous probability function $f(x)$ assume that you with to generate numbers distributed as $f(x)$ using a random number generator. If the random number generator yields a uniformly distributed value $y_i \in [0,1]$ for each trial i, the formula connecting y_i with a variable distributed as $f(x)$ is then $x_i = F^{-1}(y_i)$, where $F^{-1}(x)$ is inverse function of distribution function $F(x)$.

Probability Function

The **probability function** $f(x)$, also called **probability density function** (PDF) or **density function**, of a continuous distribution is defined as the derivative of the (cumulative) distribution function $F(x)$.

For continuous distribution we have

$$f(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \frac{dF(x)}{dx}$$

and the distribution function $F(x)$ is therefore related to a continuous probability function $f(x)$ by

$$F(x) = \mathrm{P}(X \le x) \equiv \int_{-\infty}^{x} f(x)dx$$

For discrete distribution

$$X = \begin{Bmatrix} x_1 & x_2 & \dots & x_N \\ p_1 & p_2 & & p_N \end{Bmatrix}$$

$$p_i = \mathrm{P}(X = x_i)$$

we have

$$\mathrm{P}(X = x_i) = F(x_i + 0) - F(x_i)$$

and the distribution function $F(x)$ is related to a discrete probability function $f(x)$ by

$$F(x) = \mathrm{P}(X \le x) \equiv \sum_{X \le x} f(x) = \sum_{x_i \le x} p_i$$

A probability function satisfies

$$F(x) = \mathrm{P}(x \in B) \equiv \int_B f(x)dx \qquad \text{– for continuous distribution}$$

$$F(x) = \mathrm{P}(x \in B) \equiv \sum_{x \in B} \mathrm{P}(X = x) \qquad \text{– for discrete distribution}$$

and is constrained by the normalization condition

$$\mathrm{P}(-\infty < x < \infty) = \int_{-\infty}^{\infty} f(x)dx \equiv 1 \qquad \text{– for continuous distribution}$$

$$\mathrm{P}(-\infty < x < \infty) = \sum_{i=1}^{N} p_i = 1 \qquad \text{– for discrete distribution}$$

Special case is

$$\mathrm{P}(a \le X \le b) = \int_a^b f(x)dx = F(x)\Big|_a^b = F(b) - F(a)$$

A probability function is always non-negative

$$f(x) \ge 0$$

which is equivalent to statement that distribution function $F(x)$ is monotonic non-declining function, i.e.

$$(\forall \delta > 0)\ F(x) \le F(x + \delta)$$

Quantile

The quantile is the inverse of the distribution function $F(x)$. The q$^{\text{th}}$ quantile is the value of x_q at which distribution function $F(x)$ reaches q, i.e.

$$x_q = F^{-1}(q)$$

Mode

The mode is the value of x_m at which probability function $f(x)$ reaches maximum, i.e.

$$x_m = f^{-1}(\max_{-\infty < x < \infty} f(x)).$$

Expectation Value and Population Mean

$$\mu = \langle X \rangle = \bar{X}$$

$$\langle X \rangle = \int_{-\infty}^{\infty} x f(x)dx$$

$$\langle X \rangle = \sum_{i=1}^{N} x_i p_i = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Let $Y = \varphi(X)$, then

$$\langle Y \rangle = \langle \varphi(X) \rangle = \int_{-\infty}^{\infty} \varphi(x) f(x)dx$$

Moment

The **n$^{\text{th}}$ raw moment** μ'_n, i.e. moment about zero, of a distribution $f(x)$ is defined by

$\mu'_n = \langle X^n \rangle$, i.e.

$$\mu_n' = \int_{-\infty}^{\infty} x^n f(x)dx \qquad \text{– for continuous distribution}$$

$$\mu_n' = \sum_{i=1}^{N} x_i^n p_i = \frac{1}{N}\sum_{i=1}^{N} x_i^n \qquad \text{– for discrete distribution}$$

Raw moment is sometimes called also “crude” moment.

The **nth central moment** μ_n, i.e. moment about mean, of a distribution $f(x)$ is defined by

$$\mu_n = \left\langle (X-\mu)^n \right\rangle, \text{ i.e.}$$

$$\mu_n = \int_{-\infty}^{\infty} (x-\mu)^n f(x)dx \qquad \text{– for continuous distribution}$$

$$\mu_n = \sum_{i=1}^{N} (x_i-\mu)^n p_i = \frac{1}{N}\sum_{i=1}^{N} (x_i-\mu)^n \qquad \text{– for discrete distribution}$$

Characteristic Function

The term characteristic function is denoted $\phi(t)$ and is defined as the Fourier transform of the probability density function using Fourier transform parameters $(a,b)=(0,1)$,

$$\phi(t) = \mathfrak{F}_x\left[f(x)\right](t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx$$

which is equivalent to

$$\phi(t) = \left\langle e^{itx} \right\rangle = \int_{-\infty}^{\infty} e^{itx} f(x)dx$$

$$e^{itx} = \sum_{k=0}^{\infty} \frac{(itx)^k}{k!} \Rightarrow$$

$$\phi(t) = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{\infty} \frac{(itx)^k}{k!} \right) f(x)dx$$

$$= \sum_{k=0}^{\infty} \left(\int_{-\infty}^{\infty} x^k f(x)dx \right) \frac{(it)^k}{k!} = \sum_{k=0}^{\infty} \left\langle X^k \right\rangle \frac{(it)^k}{k!} = \sum_{i=0}^{\infty} \mu_i' \frac{(it)^k}{k!}$$

A statistical distribution is not uniquely specified by its moments, but is uniquely specified by its characteristic function,

$$f(t) = \mathfrak{F}_t^{-1}\left[\phi(t)\right](x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt$$

Characteristic function can therefore be used to generate raw moments,

$$\phi^{(n)}(0) \equiv \left[\frac{d^n\phi}{dt^n}\right]_{t=0} = i^n \mu_n'$$

Moment-Generation Function

Given a random variable X and a probability distribution function $f(x)$, if there exists an $h > 0$ such that

$$M(t) \equiv \left\langle e^{tx} \right\rangle$$

for $|t| < h$, then $M(t)$ is called the moment-generating function.

For a continuous distribution,

$$M(t) = \left\langle e^{tx} \right\rangle = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

$$e^{tx} = \sum_{k=0}^{\infty} \frac{(tx)^k}{k!} \Rightarrow$$

$$M(t) = \int_{-\infty}^{\infty} \left(\sum_{k=0}^{\infty} \frac{(tx)^k}{k!} \right) f(x)dx$$

$$= \sum_{k=0}^{\infty} \left(\int_{-\infty}^{\infty} x^k f(x)dx \right) \frac{t^k}{k!} = \sum_{k=0}^{\infty} \left\langle X^k \right\rangle \frac{t^k}{k!} = \sum_{i=0}^{\infty} \mu_i' \frac{t^k}{k!}$$

As its name suggests, it can be used to generate raw moments,

$$M^{(n)}(0) \equiv \left[\frac{d^n M}{dt^n} \right]_{t=0} = \mu_n' = \langle X^n \rangle$$

Therefore, the mean and variance can be expressed in terms of raw moments as

$$\mu \equiv \langle X \rangle = \mu_1' = M'(0)$$

$$\sigma^2 \equiv \left\langle \left(X - \langle X \rangle \right)^2 \right\rangle = \mu_2' - (\mu_1')^2 = M''(0) - \left(M'(0) \right)^2$$

Mean Absolute Value

$$\theta = \Theta(X) = \left\langle \left| X - \langle X \rangle \right| \right\rangle$$

$$\Theta(X) = \int_{-\infty}^{\infty} \left| x - \langle X \rangle \right| f(x)dx$$

$$\Theta(X) = \sum_{i=1}^{N} \left| x_i - \langle X \rangle \right| p_i = \frac{1}{N} \sum_{i=1}^{N} \left| x_i - \langle X \rangle \right|$$

Variance

For a single variate X having a distribution $f(x)$ with known population mean $\mu = \langle X \rangle$, the population variance $\mathrm{var}(x)$, commonly written σ^2, is defined as,

$$\sigma^2 = \mathrm{var}(X) = \left\langle (X - \mu)^2 \right\rangle = \left\langle \left(X - \langle X \rangle \right)^2 \right\rangle = \left\langle X^2 \right\rangle - \langle X \rangle^2$$

For continuous distribution, it is given by

$$\mathrm{var}(X) = \int_{-\infty}^{\infty} \left(x - \langle X \rangle \right)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} x f(x)dx \right)^2$$

For discrete distribution with N possible values of x_i, the population variance is therefore

$$\mathrm{var}(X) = \sum_{i=1}^{N} \left(x_i - \langle X \rangle \right)^2 p_i = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \langle X \rangle \right)^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right)^2$$

Standard Deviation

$$\sigma = \sqrt{\mathrm{var}(X)}$$

Variation Coefficient

$$k_v = \frac{\sigma}{\mu} = \frac{\sqrt{\mathrm{var}(X)}}{\langle X \rangle}$$

Chebyshev Inequality

If a random variable has a finite mean $\mu = \langle X \rangle$ and finite variance $\sigma^2 = \mathrm{var}(X)$, then for each $0 < \varepsilon < 1$,

$$\mathrm{P}(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

or for each $k > 0$

$$\mathrm{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Gamma Distribution

Consider a distribution function $F(x)$ of waiting times until the p^{th} Poisson event given a Poisson distribution with a rate of change λ,

$$\begin{aligned}F(x) &= \mathrm{P}(X \le x) = 1 - \mathrm{P}(X > x)\\ &= 1 - \sum_{k=0}^{p-1} \frac{(\lambda x)^k e^{-\lambda x}}{k!} e^{-\lambda x} = 1 - e^{-\lambda x} \sum_{k=0}^{p-1} \frac{(\lambda x)^k}{k!}\\ &= 1 - \frac{\Gamma(p, \lambda x)}{\Gamma(p)}\end{aligned}$$

for $x \in [0, \infty)$, where $\Gamma(x)$ is a complete gamma function, and $\Gamma(p, x)$ an incomplete gamma function. With p an integer, the distribution is a special case known as the Erlang distribution.

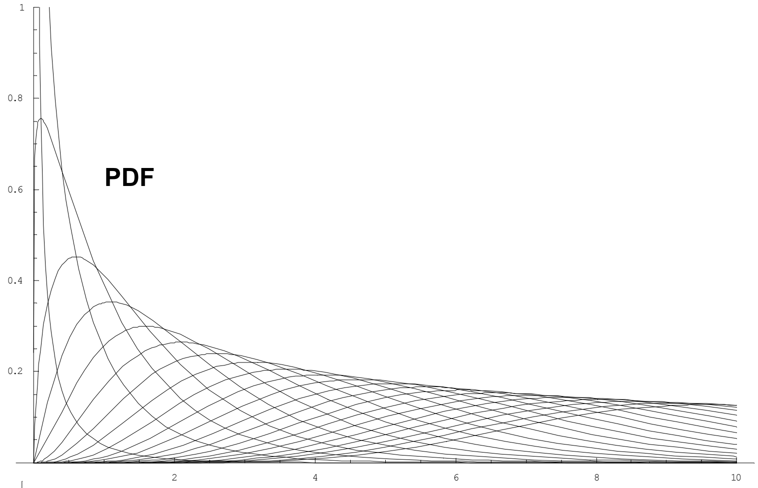
The corresponding probability function $f(x)$ of waiting times until the p^{th} Poisson event is then obtained by differentiating $F(x)$

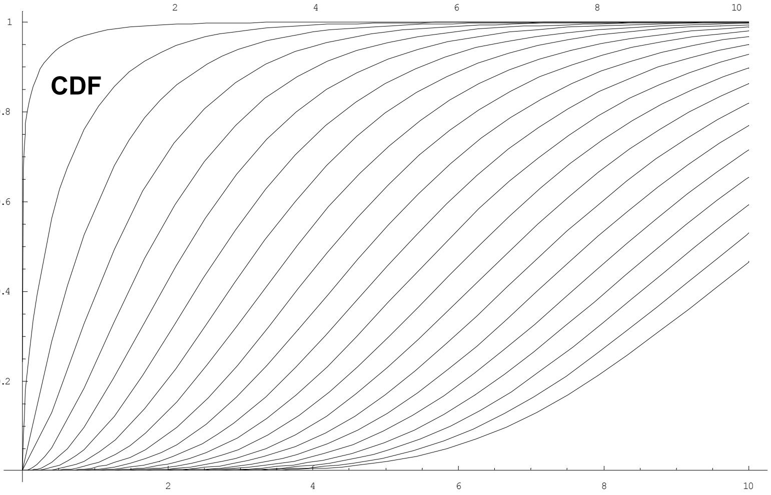
$$f(x) = F'(x) = \frac{\lambda(\lambda x)^{p-1}}{\Gamma(p)} e^{-\lambda x}$$

Parameter p is called the shape, and $\theta = {}^{1}\!/_{\lambda}$ is time between changes.

Special cases:

- Exponential Distribution: $p = 1$
- Erlang's Distribution: $p = k + 1, k \in \mathbb{N} \Rightarrow \Gamma(p-1) = k!$
- χ^2 Distribution: $p = {}^{n}\!/_{2}, n \in \mathbb{N}$





Moment-Generation Function

$$M(t) = \frac{1}{\left(1 - {}^{t}\!/_{\lambda}\right)^{p}}$$

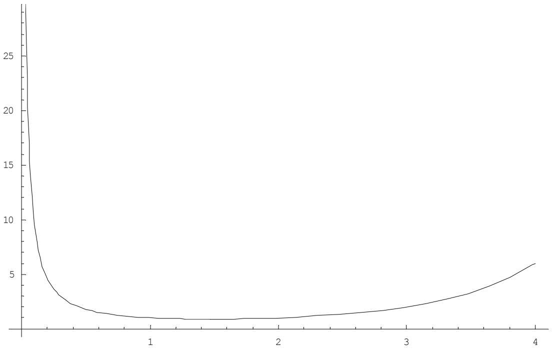
Expectation Value and Variance

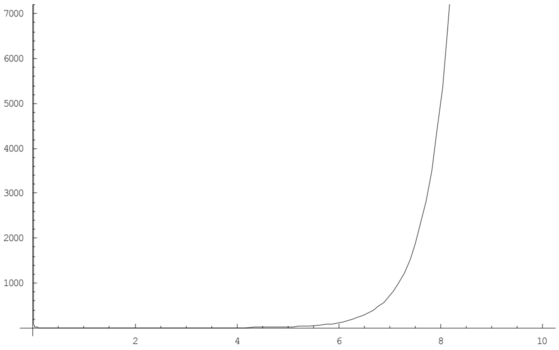
$$\mu = \frac{p}{\lambda},\ \sigma^2 = \frac{p}{\lambda^2}$$

Gamma Function

$$\Gamma(p) \equiv \int_0^\infty t^{p-1} e^{-t}\, dt$$

Note: $\left(\forall n \in \mathbb{N}_0\right)\, n! = \Gamma(n+1)$

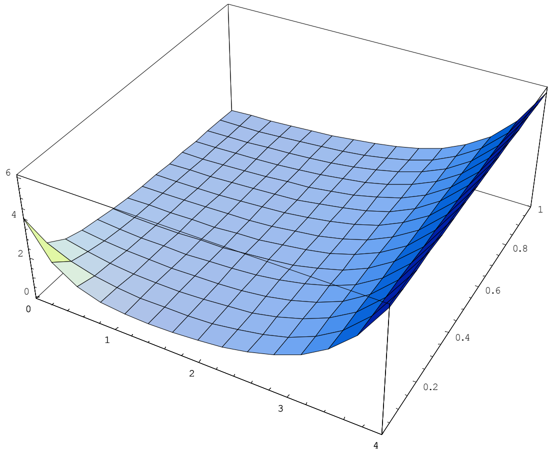




Incomplete Gamma Function

$$\Gamma(p, x) \equiv \int_0^x t^{p-1} e^{-t}\, dt$$

$$\Gamma(p) = \Gamma(p, \infty)$$



Regularized Incomplete Gamma Function

$$Q(p, x) \equiv \frac{\Gamma(p, x)}{\Gamma(p)}$$

Algorithms

Sample Population Mean and Variance Computation

When computing the sample variances numerically, the mean must be computed before σ^2 can be determined. This requires storing the set of sample values. However, it is possible to calculate σ^2 using a recursion relationship involving only the last sample as follows. This means μ itself need not be recomputed, and only a running set of values need be stored at each step.

In the following, we use somewhat less than optimal notation μ_n to denote μ calculated from the first n samples, i.e., not the n^{th} moment,

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n-1} x_i + \frac{1}{n}x_i = \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n-1} x_i\right) + \frac{1}{n}x_i = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}x_i$$

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_n\right)^2 = r_n^2 - \mu_n^2,$$

where

$$r_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{i=1}^{n-1} x_i^2 + \frac{1}{n}x_n^2 = \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=1}^{n-1} x_i^2\right) + \frac{1}{n}x_n^2 = \frac{n-1}{n}r_{n-1}^2 + \frac{1}{n}x_n^2$$

giving finally

$$\mu_n = \left(1 - \frac{1}{n}\right)\mu_{n-1} + \frac{1}{n}x_n$$

$$r_n^2 = \left(1 - \frac{1}{n}\right)r_{n-1}^2 + \frac{1}{n}x_n^2$$

$$\sigma_n^2 = r_n^2 - \mu_n^2$$

with start conditions

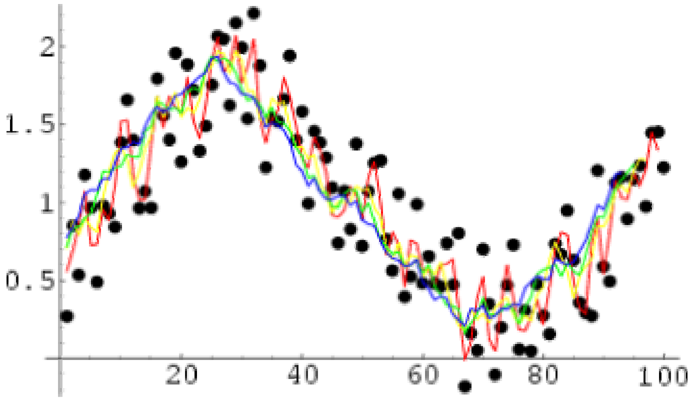
$$\mu_0 = 0,\; r_0^2 = 0$$

Moving Average

Given the sequence $\{a_i\}_{i=1}^{N-n+1}$ defined from a_i by taking average of subsequent n terms,

$$s_i = \frac{1}{n}\sum_{j=i}^{i+n-1} a_j = \frac{1}{n}\sum_{j=0}^{n-1} a_{j+i}$$

The plot bellow shows 2- (red), 4- (yellow), 6- (green) and 8- (blue) moving averages for a set of 100 data points.



Exponential Average

Averaging algorithm that is mostly used in digital signal processing and is based on the infinite impulse response low pass filter,

$$s_n = (1-\alpha)\, s_{n-1} + \alpha x_n$$

Note similarity of the sample mean and variance recursive expressions to exponential average algorithm

$$\mu_n = (1-\alpha)\, \mu_{n-1} + \alpha x_n$$

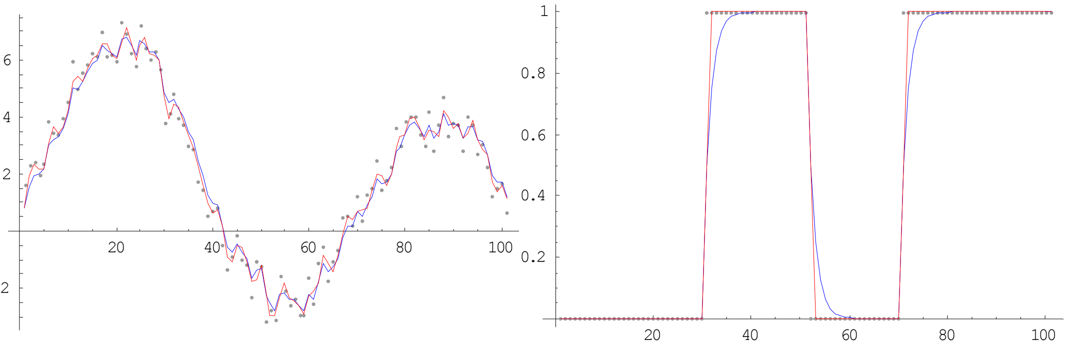
$$r_n^2 = (1-\alpha)\, r_{n-1}^2 + \alpha x_n^2$$

where α represents inverse of the τ constant of the filter, i.e. $\alpha = {}^{1}\!/_{\tau}$.

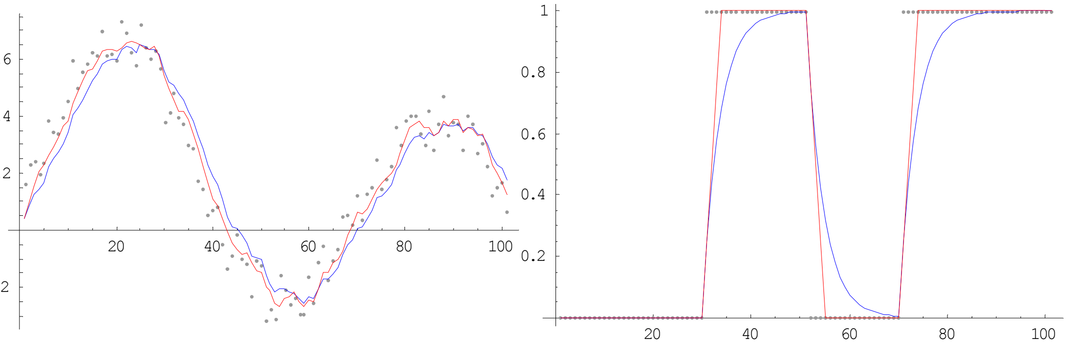
The exponential average algorithm is similar, on the other hand, to the moving average algorithm, where the size of the moving average slide window n is actually τ constant of the IIR LPF. Note, however, that IIR LPF response is exponential (and slower) and moving average response is linear (and faster).

Here is the comparison of moving average (red) and exponential average (blue) algorithms.

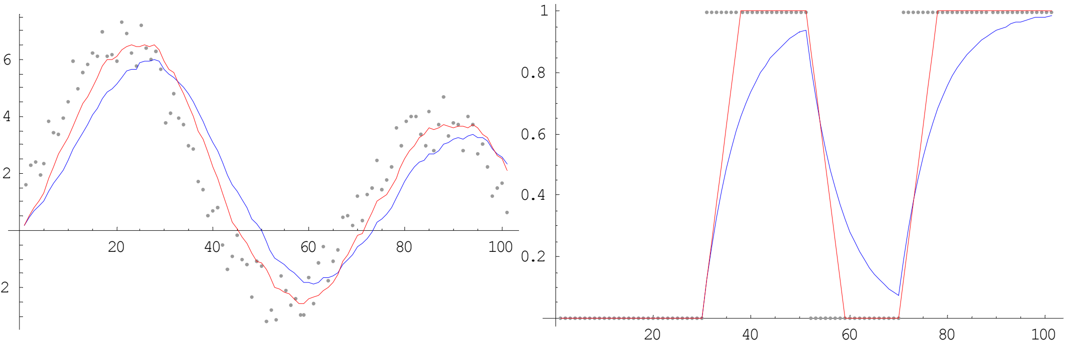
$\alpha = 0.5, n = 2$



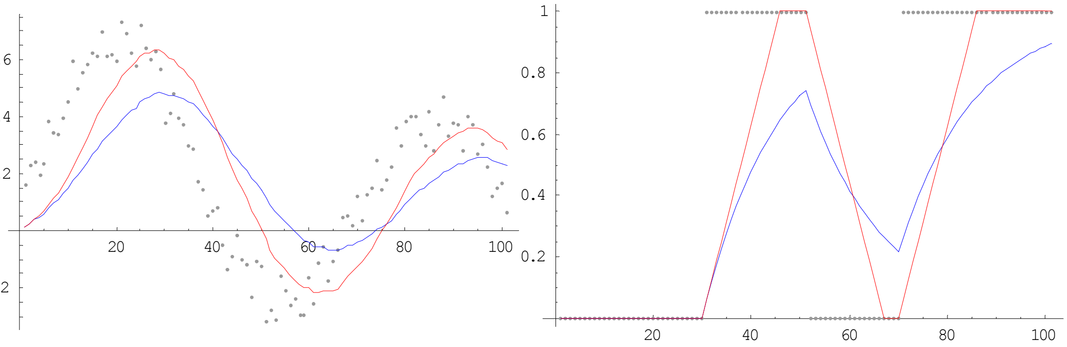
$\alpha = 0.25, n = 4$



$\alpha = 0.125, n = 8$



$\alpha = 0.0625, n = 16$


0.6
0.4
0.2
-0.2
20
40
60
80
100
1
0.8
0.6
0.4
0.2
20
40
60
80
100