Introduction to probability theory and stochastic processes

Cameron Ray Smith[1]

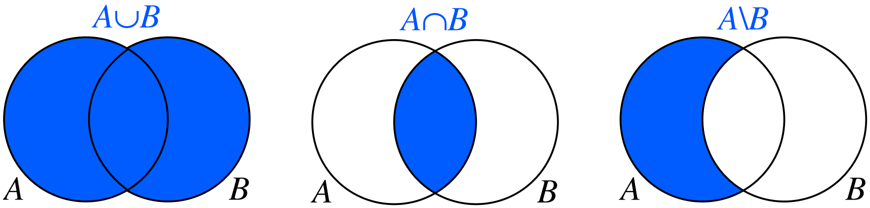
[1]Department of Systems and Computational Biology
[2]Dominick P. Purpura Department of Neuroscience
[3]Department of Pathology
Albert Einstein College of Medicine
[4]Santa Fe Institute

Albert Einstein College of Medicine

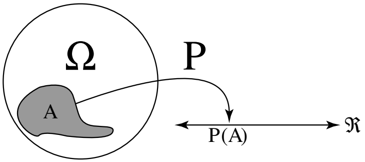
December 13, 2013

Outline

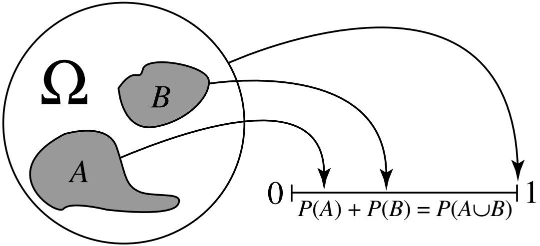
sets



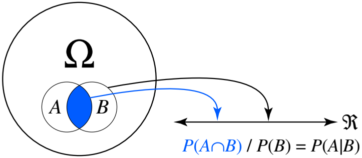
probability measure



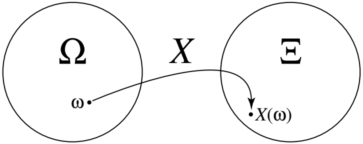
probability axioms



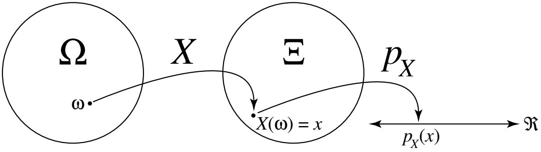
conditional probability



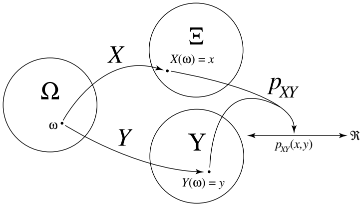
random variables



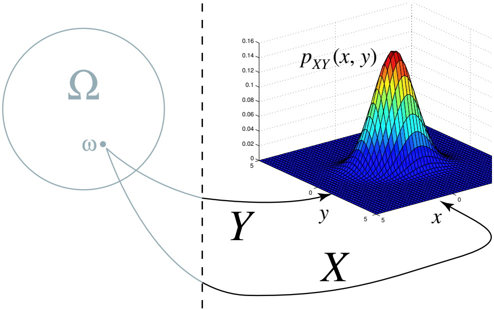
probability densities



joint probability densities



the reality



- For the general theory of measure spaces, we first need a *measurable space* (Ω, Σ), that is a set equipped with a collection Σ of **measurable sets** complete under certain operations. Then this becomes a measure space (Ω, Σ, μ) by throwing in a function μ from Σ to a space of values (such as the real line) that gets along with the set-theoretic operations that Σ has. If E is a measurable set, then $\mu(E)$ is called the **measure** of E with respect to μ. [1, 2]

1. Given a set Ω,
2. a σ-**algebra** is a collection of subsets of Ω that is closed under complementation, countable unions, and countable intersections.
3. A **measurable space**, by the usual modern definition, is a set Ω equipped with a σ-algebra Σ.
4. The elements of Σ are called the **measurable sets** of Ω (or more properly, the measurable subsets of (Ω, Σ)).

A **measure space** is a **measurable space** equipped with a **measure**. There are many different types of measures parametrized by the type of their codomains. Let (Ω, Σ) be a measurable space. A **probability measure** on Ω (due to Kolmogorov) is a function μ from the collection Σ of measurable sets to the unit interval $[0, 1]$ such that:

1. The measure of the empty set is zero: $\mu(\emptyset) = 0$;
2. The measure of the entire space is one: $\mu(\Omega) = 1$;
3. Countable additivity: $\mu(\bigcup_{i=1}^{\infty} S_i) = \sum_{i=1}^{\infty} \mu(S_i)$ whenever the S_i are mutually disjoint sets—disjoint. (Part of the latter condition is the requirement that the sum on the right-hand side must converge.)

It is sometimes stated (but in fact follows from the previous) that:

- Finitary additivity: $\mu(S \cup T) = \mu(S) + \mu(T)$ whenever S and T are disjoint.
- μ is increasing: $\mu(A) \leq \mu(B)$ if $A \subseteq B$.

Measures can be thought of in terms of integrals and densities are defined in terms of measures. Let A be, for example, one of the measurable sets from the collection of measurable sets, Σ, of our sample space Ω.

- $\mu(A) = \int_A dx$ or $\mu(A) = \int_A p(x)dx$
- $\mu(A)$ represents the mass of A which can be interpreted geometrically as an *abstract volume* or probabilistically as *the probability mass of the event "random variable X takes a value within A"*
- A **density** can then be defined intuitively as a function that transforms some measure μ_1 into a measure μ_2 by pointwise reweighting on the sample space Ω. Thus, densities are always relative measures.
- $d\mu_2(x) = f(x)d\mu_1(x)$ or $\frac{d\mu_2}{d\mu_2}(x) = f(x)$

Does a density always exist?

- A density function f is thus a function that is integrated to obtain information in terms of measure μ_2 from information in terms of measure μ_1.
- $\mu_2(A) = \int_A d\mu_2(x) = \int_A f(x) d\mu_1(x)$ is not defined if $\mu_1(A) = 0$ and $\mu_2(A) \neq 0$.
- If this is never the case for all $A \in \Sigma$, then μ_2 is referred to as *absolutely continuous* with respect to μ_1 and this relationship is written $\mu_2 \ll \mu_1$.
- This conclusion is formalized in the **Radon-Nikodym theorem** which states that μ_2 has a density with respect to μ_1 if and only if $\mu_2 \ll \mu_1$.

... so the answer is ... no, which is the reason for going through all this abstract stuff

Outline

Intuitively, **stochastic processes** are ∞-dimensional probability distributions.

- In most applications, stochastic processes model systems that evolve randomly in time, which is likely the origin of the word *process* in the name, but stochastic processes are not restricted to the temporal metaphor. Think about *order* as a generalization of time.
- The order of this evolution can be described through the use of an index and an index set respectively $t \in T$.

Definition

Consider a random experiment with sample space Ω, a σ-algebra Σ, a base probability measure $\mu : \Sigma \to [0, 1]$, and a collection of random variables X_t indexed by a set T. A **stochastic process** *is then defined by the set $\{X_t, t \in T\}$.*

- This definition can be specialized to the case of discrete or continuous stochastic processes by taking the index set to be $T \in \mathbb{N}$ or $T \in \mathbb{R}_+$ respectively.

In the discrete case imagine indexing by natural numbers, $\mathbb{N}$, such that the process could be represented as a sequence $\{X_n, n = 0, 1, 2, \ldots\}$.

How should we understand the continuous case?

- For each $\omega \in \Omega$, consider $X_t(\omega) = g_\omega(t)$. $g_\omega(t)$ can then be thought of as a function of t that realizes or samples from the stochastic process.
- For any given t, X_t is a random variable, thus to completely describe the stochastic process we need a description of the joint family of random variables $\{X_t, t \in T\}$ as opposed to just the individual random variables as if they were independent.
- For any discrete subset of times $\{t_1, \dots, t_n\}$ such that $t_1 < \dots < t_n$ and associated $\{x_1, \dots, x_n\}$ we must determine $P(X_{t_1} \leq x_1, \dots, X_{t_n} \leq x_n)$

How should we understand the continuous case?

- The **Kolmogorov extension theorem** ensures that the potentially infinite distribution $\{X_t, t \in T\}$ where, for example $T \in \mathbb{R}_+$, i.e. $T = [0, \infty)$, can be generated from the finite-dimensional families defined by $P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n)$ [3]

The necessary conditions are

- Exchangeability: for all permutations π of $1, \ldots, n$ and $x_1, \ldots, x_n$, $P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n) = P(X_{t_{\pi(1)}} \leq x_{\pi(1)}, \ldots, X_{t_{\pi(n)}} \leq x_{\pi(n)})$
- Extendability: for all $x_1, \ldots, x_n$ and $t_{n+1}, \ldots, t_{n+m}$, $P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n, X_{t_{n+1}} < \infty, \ldots, X_{t_{n+m}} < \infty)$

How should we understand the continuous case?

- Given that the conditions are satisfied then there exists a probability space $(\Omega, \Sigma, \mathbb{P})$ with an associated stochastic process $X_t : T \times \Omega \to \mathbb{R}^n$ with the families $X_{t_1}, \ldots, X_{t_n}$ as finite-dimensional marginal distributions.

Gaussian (white) noise

If random variables at different points in time are mutually independent then the joint distribution is simply the product of all its marginals

$$p_2(x_1, t_1; x_2, t_2) = p_1(x_1, t_1)p_1(x_2, t_2)$$

Gaussian (white) noise

The covariance of this distribution is

$$< (X_{t_1} - \mu_1)(X_{t_2} - \mu_2) > \\ = \int dx_1 dx_2 \ (x_1 - \mu_1)(x_2 - \mu_2) p_2(x_1, t_1; x_2, t_2) = 0 \quad (1)$$

where the mean value at time t_i is

$$\mu_i =< X_i >= \int dx \ x \, p_1(x, t_i)$$

Gaussian (white) noise

If $p_1(x,t)$ is independent of time such that $p_1(x,t) \equiv p_1(x)$ then the associated sequence of random variables is said to be identically and independently distributed (IID) with mean μ and variance σ^2

$$X(t) \sim IID(\mu, \sigma^2)$$

If $p_1(x)$ is the density of a normal distribution then this stochastic process is referred to as Gaussian noise or white noise. The $X(t)$ are then independently and normally distributed at each point in time, which we denote as

$$X(t) \sim N(0, \sigma^2)$$

Gaussian (white) noise

If we denote Gaussian noise as $\eta(t)$ then $< \eta(t) >= 0$. The covariance function between two timepoints in the discrete and continuous cases respectively are

$$
\begin{aligned}
< \eta(t)\eta(t') > &= \sigma^2 \delta_{tt'}, &(2)\\
< \eta(t)\eta(t') > &= \sigma^2 \delta(t - t'). &(3)
\end{aligned}
$$

The Fourier transform of the stationary two-time covariance function for a continuous-time process is

$$F(\omega) = \int d\tau < \eta(t)\eta(t+\tau) > e^{i\omega\tau} = \sigma^2.$$

Markov processes

first order Markov processes are characterized by the so-called Markov condition

$$p_n(x_n, t_n | x_{n-1}, t_{n-1}; \ldots ; x_1, t_1) = p_2(x_n, t_n | x_{n-1}, t_{n-1}).$$

The Markov condition means that the probability for a transition from x_{n-1} to x_n during the time interval from t_{n-1} to t_n is independent of the x_i associated to timepoints prior to t_{n-1}. All the information that is useful for attempting to make predictions is embodied in the present.

Markov processes

The conditional probability $p_2(x, t|x', t')$ is named the transition probability. A Markov process whose transition probability depends only on the time interval is called homogeneous. If, additionally, $p_1(x, t)$ is independent of t, then the process is called stationary. A Markov process is uniquely determined through $p_1(x, t)$ and $p_2(x_2, t_2|x_1, t_1)$. This is a result of the fact that the Markov property ensures that all joint distributions p_n for $n > 2$ can be expressed in terms of p_1 and p_2 as

$$\begin{aligned} &p_n(x_n, t_n; x_{n-1}, t_{n-1}; \ldots; x_1, t_1) \\ &= p_2(x_n, t_n|x_{n-1}, t_{n-1}) p_2(x_{n-1}, t_{n-1}|x_{n-2}, t_{n-2}) \cdots p_1(x_1, t_1). \end{aligned} \tag{4}$$

Markov processes

m^{th} order Markov processes are characterized by a generalized form of the first-order Markov condition

$$p_n(x_n, t_n | x_{n-1}, t_{n-1}; \ldots; x_1, t_1) \\ = p_{m+1}(x_n, t_n | x_{n-1}, t_{n-1}, \ldots, x_{n-m}, t_{n-m}). \quad (5)$$

The m^{th} order Markov condition means that the probability for a transition from x_{n-1} to x_n during the time interval from t_{n-1} to t_n is independent of the x_i associated to timepoints prior to t_{n-m}. All the information that is useful for attempting to make predictions is embodied in a history that is m timesteps long.

Markov processes

The *hidden Markov process* is defined on a pair of random variables (I, X). I is taken to be a discrete random variable over some subset of natural numbers $\{1, 2, \ldots, k\} \subset \mathbb{N}$. Realizations of X are in $\mathbb{R}^m$. Assuming one exists at all, there is a joint density function $p(i, x; t)$ whose marginals are denoted

$$p(i; t) = \int d^m x \; p(i, x; t) \tag{6}$$

$$p(x; t) = \Sigma_{i=1}^{k} p(i, x; t) \tag{7}$$

and whose conditional densities are denoted

$$p(x; t | i; t) = \frac{p(i, x; t)}{p(i; t)} \tag{8}$$

$$p(i; t | x; t) = \frac{p(i, x; t)}{p(x; t)} \tag{9}$$

Markov processes

The dynamics of the hidden Markov process can be specified under the assumption of a Gaussian first-order Markov process. In this case we can express the transition probability, where $p(x, t|i, i', x'; t')$ is a Gaussian distribution and $t > t'$, as

$$p(i, x; t|i', x'; t') = p(x, t|i, i', x'; t')p(i; t|i', x'; t').$$

Markov processes: examples

The Wiener process (also called Brownian motion or random walk) has transition probability

$$p_2(x,t|x',t') = \frac{1}{\sqrt{2\pi(t-t')}} e^{-(x-x')^2/2(t-t')},$$

and $p_1(x,t)$ can be taken to be

$$p_1(x,t) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}.$$

As $t \to 0$, $p_1(x,t)$ approaches a δ function and for $t \neq 0$, $p_1(x,t)$ is a Gaussian distribution with mean 0 and variance t. This function is also the solution of the deterministic partial differential heat conduction equation in one dimension.

Markov processes: examples

The Poisson process (a special case of more general point processes) for the random variable $N(t)$ denoting the number of *events* of some type that have occurred up to time t for the discrete variable $n \subset \mathbb{N}$, $\alpha > 0$, and for $t_2 > t_1$ has transition probability

$$p_2(n_2, t_2 | n_1, t_1) = \begin{cases} \frac{[\alpha(t_2 - t_1)]^{n_2 - n_1}}{(n_2 - n_1)!} e^{-\alpha(t_2 - t_1)}, & \text{for } n_2 \geq n_1, \\ 0 & \text{for } n_2 < n_1, \end{cases}$$

and probability that n events have taken place in the interval $[0, t]$

$$p_1(n, t) = \frac{(\alpha t)^n}{n!} e^{-\alpha t}$$

A trajectory of the Poisson process $n(t)$ consists of unit jumps at random times.

Master equations

The Chapman-Kolmogorov equation is a statement of the transitivity induced by the Markov property that the transition probability of any Markov process must satisfy

$$
\begin{aligned}
p_2(x_3, t_3; x_1, t_1) &= \int dx_2 \; p_3(x_3, t_3; x_2, t_2; x_1, t_1) \\
&= \int dx_2 \; p_2(x_3, t_3 | x_2, t_2) p_2(x_2, t_2 | x_1, t_1) p_1(x_1, t_1)
\end{aligned}
$$

And the conditional probability is then

$$
p_2(x_3, t_3 | x_1, t_1) = \int dx_2 \; p_2(x_3, t_3 | x_2, t_2) p_2(x_2, t_2 | x_1, t_1)
$$

Master equations

If

$$\lim_{t\to\infty} p_1(x,t) \equiv p_{stat}(x) \neq 0$$

exists, then p_{stat} is named the *stationary distribution*, which is achieved as t approaches ∞. Now, because

$$p_1(x_3,t_3) = \int dx_2\ p_2(x_3,t_3|x_2,t_2)p_1(x_2,t_2),$$

then for the stationary distribution

$$p_{stat}(x) = \int dx_2\ p_2(x,t_3|x_2,t_2)p_{stat}(x_2),$$

for all t_2 and t_3.

Master equations

We may place some conditions on the infinitesimal-time behavior of the transition probabilities in order to use the Chapman-Kolmogorov equation to derive a differential equation specifying the dynamics of the density function characterizing the stochastic process.

Imagine a Markov process with transition probability $p_2(x, t+\tau|x'', t)$. For small (and secretly infinitesimal) values of τ the form of this transition probability will be taken to be

$$p_2(x, t+\tau|x'', t) = [1 - a(x,t)\tau]\delta(x - x'') + \tau w(x, x'', t) + O(\tau^2)$$

Master equations

Given that the transition probability $p_2(x, t+\tau|x'', t)$ is assumed to be normalized, which means that the integral over its domain of definition is 1, then the equation

$$p_2(x, t+\tau|x'', t) \\ = [1 - a(x, t)\tau]\delta(x - x'') + \tau w(x, x'', t) + O(\tau^2)$$

implies that

$$a(x'', t) = \int dx \; w(x, x'', t)$$

Master equations

From the Chapman-Kolmogorov equation we can derive

$$
\begin{aligned}
p_2(x, t+\tau|x', t') &= \int dx'' \; p_2(x, t+\tau|x'', t) p_2(x'', t|x', t') \\
&= [1 - a(x,t)\tau] p_2(x, t|x', t') \\
&\quad + \tau \int dx'' \; w(x, x'', t) p_2(x'', t|x', t') \\
&\quad + O(\tau^2)
\end{aligned}
$$

And the master equation results from taking the limit $\tau \to \infty$

$$
\begin{aligned}
\frac{\partial}{\partial t} p_2(x, t|x', t') &= \int dx'' \; w(x, x'', t) p_2(x'', t|x', t') \\
&\quad - \int dx'' \; w(x'', x, t) p_2(x, t|x', t')
\end{aligned}
$$

Master equations

A more intuitive form of the master equation given knowledge of the form of the initial distribution is obtained via multiplication by $p_1(x', t')$ and integration over x' to obtain

$$\begin{aligned} \frac{\partial}{\partial t} p_1(x,t) &= \int dx' \; w(x, x', t) p_1(x', t) \\ &- \int dx' \; w(x', x, t) p_1(x, t) \end{aligned}$$

If the variable x is taken to be discrete $n \in \mathbb{N}$, then $p_n(t) \equiv p_1(x,t)$ and $w_{nn'}(t) \equiv w(x, x', t)$ results in

$$\dot{p}_n(t) = \Sigma_{n'} [w_{nn'}(t) p_{n'}(t) - w_{n'n}(t) p_n(t)]$$

This discrete form is perhaps the most intuitive. It says that the change in probability of finding the label n at time t can be described in terms of two components. The first is the number of transitions into (or creations) $n' \to n$ and the second is the number of transitions from (or annihilations) $n \to n'$.

Master equations

$$\dot{p}_n(t) = \Sigma_{n'}[w_{nn'}(t)p_{n'}(t) - w_{n'n}(t)p_n(t)]$$

This discrete state form is perhaps the most intuitive so far. It says that the change in probability of finding the label n at time t can be described in terms of two components. The first is the number of transitions into (or creations) $n' \to n$ and the second is the number of transitions from (or annihilations) $n \to n'$.

The Fokker-Planck Equation

Stochastic differential equations

Outline

Population genetics

TODO: topics to cover

Determinism and stochasticity in gene expression

- Elowitz [4]
- Paulsson [5, 6, 7]
- Van Oudenaarden [8, 9]

Outline

[1] David Insua, Fabrizio Ruggeri, and Mike Wiper.
Bayesian Analysis of Stochastic Process Models.
Wiley, 2012.

[2] NLab.
Measure space.
http://ncatlab.org/nlab/show/measure+space.

[3] Bernt Ø ksendal.
Stochastic Differential Equations: An Introduction with Applications.
Springer, 2010.

[4] Peter S Swain, Michael B Elowitz, and Eric D Siggia.
Intrinsic and extrinsic contributions to stochasticity in gene expression.
PNAS, 99(20):12795–800, October 2002.

[5] Johan Paulsson.
Summing up the noise in gene networks.
Nature, 427(6973):415–8, January 2004.

[6] Andreas Hilfinger and Johan Paulsson.
Separating intrinsic from extrinsic fluctuations in dynamic biological systems.
PNAS, 2011(9), July 2011.

[7] Ioannis Lestas, Glenn Vinnicombe, and Johan Paulsson.
Fundamental limits on the suppression of molecular fluctuations.
Nature, 467(7312):174–178, September 2010.

[8] Mukund Thattai and Alexander van Oudenaarden.
Stochastic gene expression in fluctuating environments.
Genetics, 167(1):523–30, May 2004.

[9] B. Munsky, G. Neuert, and A. van Oudenaarden.
Using Gene Expression Noise to Understand Gene Regulation.
Science, 336(6078):183–187, April 2012.