Stochastic Calculus Notes, Lecture 4

Last modified March 2, 2007

1 Continuous probability

1.1. Introduction: Recall that a set Ω is *discrete* if it is finite or countable. We will call a set *continuous* if it is not discrete. Many of the probability spaces used in stochastic calculus are continuous in this sense (examples below). Kolmogorov[1] suggested a general framework for continuous probability based on abstract integration with respect to abstract probability measures. The theory makes it possible to discuss general constructions such as conditional expectation in a way that applies to a remarkably diverse set of examples.

The difference between continuous and discrete probability is the difference between integration and summation. Continuous probability cannot be based on the formula

$$P(A) = \sum_{\omega \in A} P(\omega) \; . \tag{1}$$

Indeed, the typical situation in continuous probability is that any single outcome has probability zero: $P(\{\omega\}) = 0$ for all $\omega \in \Omega$.

As we explain below, the classical formalism of probability densities also does not apply in many of the situations we are interested in. Abstract probability measures give a framework for working with probability in path space, as well as more traditional discrete probability and probabilities given by densities on R^n.

These notes outline the Kolmogorov's formalism of probability measures for continuous probability. We leave out a great number of details and mathematical proofs. Attention to all these details would be impossible within our time constraints. In some cases we indicate where a precise definition or a complete proof is missing, but sometimes we just leave it out. If it seems like something is missing, it may be.

1.2. Examples of continuous probability spaces: A *probability space* is a set, Ω, of possible outcomes, together with a $\sigma-$algebra, $\mathcal{F}$, of measurable events. This section discusses only the sets Ω. The corresponding algebras are discussed below.

R, the real numbers. If x_0 is a real number and $u(x)$ is a probability density, then the probability of the event $B_r(x_0) = \{x_0 - r \leq X \leq x_0 + r\}$ is

$$P([x_0 - r, x_0 + r]) = \int_{x_0-r}^{x_0+r} u(x)dx \to 0 \;\text{ as } r \to 0.$$

[1]The Russian mathematician Kolmogorov was active in the middle of the 20^{th} century. Among his many lasting contributions to mathematics are the modern axioms of probability and some of its most important theorems. His theories of turbulent fluid flow anticipated modern fractals be several decades.

Thus the probability of any individual outcome is zero. An event with positive probability ($P(A) > 0$) is made up entirely of outcomes $x_0 \in A$, with $P(x_0) = 0$. Because of countable additivity (see below), this is only possible when Ω is uncountable.

R^n, sequences of n numbers (possibly viewed as a row or column vector depending on the context): $X = (X_1 \ldots, X_n)$. Here too if there is a probability density then the probability of any given outcome is zero.

$\mathcal{S}^{\mathcal{N}}$. Let $\mathcal{S}$ be the discrete state space of a Markov chain. The space $\mathcal{S}^T$ is the set of sequences of length T of elements of $\mathcal{S}$. An element of $\mathcal{S}^T$ may be written $x = (x(0), x(1), \cdots, x(T-1))$, with each of the $x(t)$ in $\mathcal{S}$. It is common to write x_t for $x(t)$. An element of $\mathcal{S}^{\mathcal{N}}$ is an infinite sequence of elements of $\mathcal{S}$. The "exponent" $\mathcal{N}$ stands for "natural numbers". We misuse this notation because ours start with $t = 0$ while the actual natural numbers start with $t = 1$. We use $\mathcal{S}^{\mathcal{N}}$ when we ask questions about an entire infinite trajectory. For example the hitting probability is $P(X(t) = 1 \text{ for some } t \geq 0)$. Cantor proved that $\mathcal{S}^{\mathcal{N}}$ is not countable (whenever the state space has more than one element). In most cases, the probability of any particular infinite sequence is zero. For example, suppose the transition matrix has $P_{11} = .6$ and $X(0) = 1$ (so $u_0(1) = 1$). Let x be the infinite sequence that never leaves state 1: $x = (1, 1, 1, \cdots)$. Then $P(x) = u_0(1) \cdot .6 \cdot .6 \cdots$. Multiplying together an infinite number of .6 factors should give the answer $P(x) = 0$. More generally, if the transition matrix has $P_{jk} \leq r < 1$ for all (j, k), then $P(x) = 0$ for any single infinite path.

$C([0, T] \to R)$, the path space for Brownian motion. The C stands for "continuous". The $[0, T]$ is the time interval $0 \leq t \leq T$; the square brackets tell us to include the endpoints (0 and T in this case). Round parentheses $(0, T)$ would mean to leave out 0 and T. The final R is the "target" space, the real numbers in this case. An element of Ω is a continuous function from the interval $[0, T]$ to R. This function could be called $X(t)$ or X_t (for $0 \leq t \leq T$). In this space we can ask questions such as $P(\int_0^T X(t)dt > 4)$.

1.3. Probability measures: Let $\mathcal{F}$ be a $\sigma-$algebra of subsets of Ω. A *probability measure* is a way to assign a probability to each event $A \in \mathcal{F}$. In discrete probability, this is done using (1). In R^n a probability density leads to a probability measure by integration

$$P(A) = \int_A u(x)dx \ . \tag{2}$$

There are still other ways to specify probabilities of events in path space. All of these probability measures satisfy the same basic axioms.

Suppose that for each $A \in \mathcal{F}$ we have a number $P(A)$. The numbers $P(A)$ form a *probability measure* if

i. If $A \in \mathcal{F}$ and $B \in \mathcal{F}$ are disjoint events, then $P(A \cup B) = P(A) + P(B)$.

ii. $P(A) \geq 0$ for any event $A \in \mathcal{F}$.

iii. $P(\Omega) = 1$.

iv. If $A_n \in \mathcal{F}$ is a sequence of events each disjoint from all the others and $\cup_{n=1}^{\infty} A_n = A$, then $\sum_{n=1}^{\infty} P(A_n) = P(A)$.

The last property is called *countable additivity*. It is possible to consider probability measures that are not countably additive, but is not bery useful. Other properties of probabilities follow from these. For example, if A^c is the event that A did not happen, then $P(A^c) = 1 - P(A)$ since A and A^c are disjoint events and $A \cup A^c = \Omega$.

1.4. Example, discrete probability: If Ω is discrete, we may take $\mathcal{F}$ to be the set of all events (i.e. all subsets of Ω). If we know the probabilities of each individual outcome, then the formula (1) defines a probability measure. The axioms (i), (ii), and (iii) are clear. The last, countable additivity, can be verified given a solid undergraduate analysis course.

1.5. Generating a probability measure: It is rare in continuous probability that one can define $P(A)$ for all $A \subseteq \Omega$. Usually, there are *non measurable* events whose probability one does not try to define (see below). This is not related to partial information, but is an intrinsic aspect of continuous probability. Instead, we find a class of *basic events*, A, for which $P(A)$ is easy to define. We then take $\mathcal{F}$ to be the $\sigma-$algebra these events generate. The *Kolmogorov extension theorem*, which we do not even state, then tells us (in favorable cases) that $P(A)$ also is defined for any $A \in \mathcal{F}$. Non measurable events are artificial in the sense that they cannot be represented as limits of basic events.

1.6. Borel sets: In most applications in stochastic calculus, the basic sets are *balls*. Events in the $\sigma-$algebra generated by balls are called *Borel sets*.[2] In R^n, the *open ball* with center x_0 and radius $r > 0$ is $B_r(x_0) = \{x \text{ with } |x - x_0| < r\}$. A "ball" in one dimension is an interval. In two dimensions it is a disk. Note that the ball is solid, as opposed to the hollow *sphere*, $S_r(x_0) = \{x \text{ with } |x - x_0| = r\}$. The condition $|x - x_0| \leq r$, instead of $|x - x_0| \leq r$, defines a *closed* ball. The $\sigma-$algebra generated by open balls is the same as that generated by closed balls (check this if you wish). You also can show (if you wish) that the $\sigma-$ algebra generated by balls contains, for example, rectangles, smooth curves, etc.

1.7. Borel sets in path space: The definition of Borel sets works the same way in the path space of Brownian motion, $C([0,T], R)$. Let $x_0(t)$ and $x(t)$ be

[2]The larger $\sigma-$algebra of *Lebesgue sets* is more of a nuisance than a help, because the question of which events are measurable depends on which probability measure is being used. It is hard to compare different probability measures when even the $\sigma-$algebras are different.

two continuous functions of t. The distance between them in the *sup norm* is

$$\|x - x_0\| = \sup_{0 \le t \le T} |x(t) - x_0(t)| \ .$$

We often use double bars to represent the distance between functions and single bar absolute value signs to represent the distance between numbers or vectors in R^n. As before, the open ball of radius r about a path x_0 is the set of all paths with $\|x - x_0\| < r$.

1.8. The $\sigma-$algebra for Markov chain path space: There is a convenient limit process that defines a useful $\sigma-$algebra on $\mathcal{S}^{\mathcal{N}}$, the infinite time horizon path space for a Markov chain. We have the algebras $\mathcal{F}_T$ generated by the first $T+1$ states $x(0)$, $x(1)$, $\ldots$, $x(T)$. We take $\mathcal{F}$ to be the $\sigma-$algebra generated by all these. Note that the event $A = \{X(t) \neq 1 \text{ for } t \ge 0\}$ is not in any of the $\mathcal{F}_T$. However, the event $A_T = \{X(t) \neq 1 \text{ for } 0 \le t \le T\}$ is in $\mathcal{F}_T$. Therefore $A = \cap_{T \ge 0} A_T$ must be in any $\sigma-$algebra that contains all the $\mathcal{F}_T$. Also note that the union of all the $\mathcal{F}_T$ is an algebra of sets, though it is not a $\sigma-$algebra.

1.9. Non measurable sets (technical aside): A construction demonstrates that non measurable sets are unavoidable. Let Ω be the unit circle. The simplest probability measure on Ω would seem to be uniform measure (divided by 2π so that $P(\Omega) = 1$). This measure is *rotation invariant*: if A is a measurable event having probability $P(A)$ then the event $A + \theta = \{x + \theta \mid x \in A\}$ is measurable and has $P(A+\theta) = P(A)$. It is possible to construct a set B and a (countable) sequence of rotations, θ_n, so that the events $B + \theta_k$ and $B + \theta_n$ are disjoint if $k \neq n$ and $\bigcup_n (B + \theta_n) = \Omega$. This set cannot be measurable. If it were and $\mu = P(B)$ then there would be two choices: $\mu = 0$ or $\mu > 0$. In the former case we would have $P(\Omega) = \sum_n P(B + \theta_n) = \sum_n 0 = 0$, which is not what we want. In the latter case, again using countable additivity, we would get $P(\Omega) = \infty$.

The construction of the set B starts with a description of the θ_n. Write n in base ten, flip over the decimal point to get a number between 0 and 1, then multiply by 2π. For example for $n = 130$, we get $\theta_n = \theta_{130} = 2\pi \cdot .031$. Now use the θ_n to create an equivalence relation and partition of Ω by setting $x \sim y$ if $x = y + \theta_n \pmod{2\pi}$ for some n. The reader should check that this is an equivalence relation ($x \sim y \to y \sim x$, and $x \sim y$ and $y \sim z \to x \sim z$). Now, let B be a set that has exactly one representative from each of the equivalence classes in the partition. Any $x \in \Omega$ is in one of the equivalence classes, which means that there is a $y \in B$ (the representative of the x equivalence class) and an n so that $y + \theta_n = x$. That means that any $x \in \Omega$ has $x \in B + \theta_n$ for some n, which is to say that $\bigcup_n (B + \theta_n) = \Omega$. To see that $B + \theta_k$ is disjoint from $B + \theta_n$ when $k \neq n$, suppose that $x \in B + \theta_k$ and $x \in B + \theta_n$. Then $x = y + \theta_k$ and $x = z + \theta_n$ for $y \in B$ and $z \in B$. But (and this is the punch line) this would mean $y \sim z$, which is impossible because B has only one representative from each equivalence class.

The possibility of selecting a single element from each partition element without having to say how it is to be done is the *axiom of choice*. It is impossible to create a non-measureable set without the axiom of choice (this is a deep and obscure theorem in set theory). In practical applications of probability it might happen that an event is not in some $\sigma-$algebra because there is not enough information. It never happens that an event is not measureable because of the existence of non-measurable sets as above.

1.10. Probability measures in R^n: Suppose $u(x)$ is a probability density in R^n. If A is an event made from finitely many balls by set operations, we can define $P(A)$ by integrating, as in (2). This leads to a probability measure on Borel sets corresponding to the density u.

1.11. Singular probability measures in R^n: There are many interesting probability measures in R^n that do not have proper probability densities. For example, if the "random" variable, X takes only the value x_0, the probability measure is called a *point mass*, or *delta function*[3] If A is a Borel set, the point mass measure is defined by $P_{x_0}(A) = 1$ if $x_0 \in A$, and $P_{x_0} = 0$ if $x_0 \notin A$.

There also are *mixtures* of point masses. Suppose we have points $x_k \in R^n$, and probability weights $q_k > 0$ (and $\sum_k q_k = 1$), there the measure that gives X probability q_k to be at x_k is $\sum_k q_k P_{k_k}$. It is given by $P(A) = \sum k \in A q_k$.

Besides being *concentrated* on discrete points as above, a probability measure in the plane ($\Omega = R^2$) can be concentrated on a curve, C. If $A \subseteq R^2$ is disjoint from C then $P(A) = 0$. For example, suppose $U \in R$ is a univariate random variable with uniform probability density and $X = (X_1, X_2) \in R^2$ is given by $X_1 = \cos(U)$, $X_2 = \sin(U)$. This defines a probability measure concentrated on the unit circle in the plane.

1.12. Measurable functions: Let Ω be a probability space with a $\sigma-$algebra $\mathcal{F}$. Let $f(\omega)$ be a function defined on Ω. In discrete probability, f was measurable with respect to $\mathcal{F}$ if the sets $B_a = \{\omega \mid f(\omega = a)\}$ all were measurable. In continuous probability, this definition is replaced by the condition that the sets $A_{ab} = \{\omega \mid a \leq f(\omega) \leq b\}$ are measurable. Because $\mathcal{F}$ is countably additive, and because the event $a < f$ is the (countable) union of the events $a + \frac{1}{n} \leq f$, this is the same as requiring all the sets $\widetilde{A}_{ab} = \{\omega \mid a < f(\omega) < b\}$ to be measurable. If Ω is discrete (finite or countable), then the two definitions of measurable function agree.

In continuous probability, the notion of measurability of a function with respect to a $\sigma-$algebra plays two roles. The first, which is purely technical, is that f is sufficiently "regular" (meaning not crazy) that abstract integrals (defined below) make sense for it. The second, particularly for smaller algebras $\mathcal{G} \subset \mathcal{F}$, again involves incomplete information. A function that is measurable with respect to $\mathcal{G}$ not only needs to be regular, but also must depend on fewer

[3] English Physicist Paul A. M. Dirac defined a *generalized function*, $\delta(x)$, with the property that $\int_a^b \delta(x)dx = 1$ if $a \leq 0 \leq b$ and $\int_a^b \delta(x)dx = 0$ if $b < 0$ or $a > 0$. It is clear that $\delta(x) = 0$ if $x \neq 0$. If $\delta(x)$ were an honest function, this would imply that $\int_a^b \delta(x)dx = 0$ for any a and b.

variables (possibly in some abstract sense).

1.13. Integration with respect to a measure: We want to define integration and expected value for abstract probability measures. The strategy is to list properties we want this abstract integration to have, then show that there is an operation with these properties. As with the Riemann integral of calculus, we usually do not return to the definition of the integral every time we want to evaluate one.

The definition of integration with respect to a general probability measure is easier than the definition of the Riemann integral. The integral is written

$$E[f] = \int_{\omega\in\Omega} f(\omega)dP(\omega) \; .$$

We will see that in R^n with a density u, this agrees with the classical definition

$$E[f(X)] = \int_{R^n} f(x)u(x)dx \; ,$$

if we write $dP(x) = u(x)dx$. Note that the abstract variable ω is replaced by the concrete variable, x, in this concrete situation. The general definition is forced on us once we make the natural requirements

i. If $A \in \mathcal{F}$ is any event, then $E[\mathbf{1}_A] = P(A)$. The integral of the indicator function if an event is the probability of that event.

ii. If f_1 and f_2 have $f_1(\omega) \leq f_2(\omega)$ for all $\omega \in \Omega$, then $E[f_1] \leq E[f_2]$. (Integration is monotone.)

iii. For any reasonable functions f_1 and f_2 (e.g. bounded), we have $E[af_1 + bf_2] = aE[f_1] + bE[f_2]$. (*Linearity* of integration).

1.14. Integral limit theorems: There are limit theorems for abstract integrals that are related to the property of countable additivity of $\sigma-$algebras and probability measures. Suppose $f_n(\omega)$ is a sequence of functions so that $f_n(\omega) \to f(\omega)$ for every ω as $n \to \infty$. Limit theorems are conditions under which we know that

$$\int f_n(\omega)dP(\omega) = E\left[f_n\right] \to E\left[f\right] \; . \tag{3}$$

For example, this is true the f_n are uniformly bounded (there is an M with $|f_n(\omega)| \leq M$ for every n and ω).

The limit theorems also can be expressed as countable additivity, finding conditions under which

$$\sum_{k=1}^{\infty} E\left[g_k\right] = E\left[\sum_{k=1}^{\infty} g_k\right] \; . \tag{4}$$

This is the same as (3) because we can take $f_n(\omega) = \sum_{k=1}^{n} g_k(\omega)$. For each ω, (this is the definition of an infinite sum)

$$f(\omega) = \lim_{n\to\infty} f_n(\omega) = \sum_{k=1}^{\infty} g_k(\omega) \; .$$

Also,

$$\lim_{n\to\infty} E\left[f_n\right] = \sum_{k=1}^{\infty} E\left[g_k\right] \; .$$

Just below, we will use the simple *monotone convergence theorem*. It states that (3) holds if $f_1(\omega) \geq 0$ for all ω (written simply $f_1 \geq 0$), and $f_{n+1} \geq f_n$ for all n. Part of the theorem is that if either side is infinite, the other also is infinite. This may be restated as saying that (4) holds whenever $g_k \geq 0$ for all k.

1.15. Simple functions: A function is a *simple function* if there are events A_k, and weights w_k, so that $f = \sum_k w_k 1_{A_k}$. Properties (i) and (iii) imply that the expectation of a simple function is

$$E[f] = \sum_k w_k P(A_k) \; . \tag{5}$$

The monotone convergence theorem formula (4) suggests that this should be true for infinite sums provided $w_k \geq 0$ for all k. The definition of simple function and the formula (5) do not depend on the A_k being disjoint, though they often are in specific applications.

We indicate just a few of the technicalities involved with this definition. First one should check that it depends only on f. If $f = \sum_l u_l \mathbf{1}_{B_l}$ (different weights u_l and events $B_l \in \mathcal{F}$), then

$$\sum_k w_k P(A_k) = \sum_l u_l P(B_l) \; . \tag{6}$$

To see what this means, suppose the A_k are disjoint, the w_k are equal and the single B event is $B = \cup_k A_k$. Then (6) is equivalent to $\sum_k P(A_k) = P(B)$, which is countable additivity of the probability measure P. Second is the fact that the definition (5) is additive. If $g = \sum_k v_k \mathbf{1}_{A_k}$, then

$$E\left[f+g\right] = \sum_k (w_k + v_k) P(A_k) = \sum_k w_k P(A_k) + \sum_k v_k P(A_k) \; ,$$

the last being true when the w_k and v_k are non-negative. We can assume that f and g are defined using the same events A_k because of (6) (think this through). Finally, one can verify the monotone convergence theorem for simple functions converging to a simple function. It suffices to do this when the simple function has the form $f = w\mathbf{1}_A$ (a single event in the sum).

1.16. General measurable functions: The monotonicity requirement (ii) above and the formula (5) allow us to define the abstract integral of any nonnegative measurable function. Suppose f is a nonnegative function: $0 \leq f(\omega) \leq M$ for all $\omega \in \Omega$. Choose a small number $\epsilon = 2^{-n}$, for $k \geq 0$ define the heights $h_k = \epsilon(k-1)$, and define the[4] *ring sets* $A_k = \{h_{k-1} \leq f < h_k\}$. The A_k depend on ϵ but we do not indicate that. Although the events A_k might be complicated, fractal, or whatever, each of them is measurable. A simple function that approximates f is $f_n = \sum_k h_{k-1}\mathbf{1}_{A_k}$. This f_n takes the value h_{k-1} on the sets A_k. Note that $f_n(\omega) \leq f(\omega)$ for each $\omega \in \Omega$, though by at most ϵ. Property (ii) implies that (if $E[f]$ is defined)

$$E[f] \geq E[f_n] = \sum_k h_{k-1}P(A_k) \; .$$

In the same way, we can consider the upper function $g_n = \sum_k h_k \mathbf{1}_{A_k}$, which has $g_n \geq f$ so

$$E[f] \leq E[g_n] = \sum_k h_k P(A_k) \; .$$

The reader can check that $f_n \leq f_{n+1} \leq f \leq g_{n+1} \leq g_n$ and that $g_n - f_n \leq \epsilon$ (i.e. $g_n(\omega) - f_n(\omega) \leq \epsilon$ for all ω). Therefore $E[g_n] - E[f_n] \leq \epsilon = 2^{-n}$, the numbers $E[f_n]$ form an increasing sequence, and the $E[g_n]$ are a decreasing sequence. All this implies (think this through) that the sequences $E[f_n]$ and $E[g_n]$ converge to the same number converging to the same number, which is the only possible value of $E[f]$ consistent with (i), (ii), and (iii).

It is sometimes said that the difference between classical (Riemann) integration and abstract integration (here) is that the Riemann integral cuts the x axis into little pieces, while the abstarct integral cuts the y axis (which is what the simple function approximations amount to).

1.17. Markov chain probability measures on $\mathcal{S}^{\mathcal{N}}$: Let $\mathcal{A} = \cup_{t\geq 0}\mathcal{F}_t$ as before. The probability of any $A \in \mathcal{A}$ is given by the probability of that event in $\mathcal{F}_t$ if $A \in \mathcal{F}_t$. Therefore $P(A)$ is given by a formula like (1) for any $A \in \mathcal{A}$. Let $\mathcal{F}$ be the $\sigma-$algebra generated by $\mathcal{A}$. The Kolmogorov extension theorem allows us to conclude (once we have verified its hypotheses) that there is a unique countable additive measure on $\mathcal{F}$ that agrees with $P(A)$ for $A \in \mathcal{A}$. For example, suppose $X(t)$ is simple random walk on the interval $[0, L]$ and A is the event $\{X(t) = 0 \text{ before } X(t) = L\}$. This is in $\mathcal{F}$ but not in $\mathcal{A}$.

1.18. Conditional expectation: We have a random variable $X(\omega)$ that is measurable with respect to the $\sigma-$algebra, $\mathcal{F}$. We have $\sigma-$algebra that is a sub algebra: $\mathcal{G} \subset \mathcal{F}$. We want to define the conditional expectation $Y = E[X \mid \mathcal{G}]$. In discrete probability this is done using the partition defined by $\mathcal{G}$. The partition is less useful because it probably is uncountable, and because each partition element, $B(\omega) = \cap A$ (the intersection being over all $A \in \mathcal{G}$ with

[4] Take $f = f(x,y) = x^2 + y^2$ in the plane to see why we call them ring sets.

$\omega \in A$), may have $P(B(\omega)) = 0$ (examples below). This means that we cannot apply Bayes' rule directly.

The definition is that $Y(\omega)$ is the random variable measurable with respect to $\mathcal{G}$ that best approximates X in the least squares sense

$$E[(Y - X)^2] = \min_{Z \in \mathcal{G}} E[(Z - X)^2] \; .$$

This is one of the definitions we gave before, the one that works for continuous and discrete probability. In the theory, it is possible to show that there is a minimizer and that it is unique.

1.19. Generating a $\sigma-$algebra: When the probability space, Ω, is finite, we can understand an algebra of sets by using the partition of Ω that generates the algebra. This is not possible for continuous probability spaces. Another way to specify an algebra for finite Ω was to give a function $X(\omega)$, or a collection of functions $X_k(\omega)$ that are supposed to be measurable with respect to $\mathcal{F}$. We noted that any function measurable with respect to the algebra generated by functions X_k is actually a function of the X_k. That is, if $F \in \mathcal{F}$ (abuse of notation), then there is some function $u(x_1, \ldots, x_n)$ so that

$$F(\omega) = u(X_1(\omega), \ldots, X_n(\omega)) \; . \tag{7}$$

The intuition was that $\mathcal{F}$ contains the information you get by knowing the values of the functions X_k. Any function measurable with respect to this algebra is determined by knowing the values of these functions, which is precisely what (7) says. This approach using functions is often convenient in continuous probability.

If Ω is a continuous probability space, we may again specify functions X_k that we want to be measurable. Again, these functions generate an algebra, a $\sigma-$algebra, $\mathcal{F}$. If F is measurable with respect to this algebra then there is a (Borel measurable) function $u(x_1, \ldots)$ so that $F(\omega) = u(X_1, \ldots)$, as before. In fact, it is possible to define $\mathcal{F}$ in this way. Saying that $A \in \mathcal{F}$ is the same as saying that $\mathbf{1}_A$ is measurable with respect to $\mathcal{F}$. If $u(x_1, \ldots)$ is a Borel measurable function that takes values only 0 or 1, then the function F defined by (7) defines a function that also takes only 0 or 1. The event $A = \{\omega \mid F(\omega) = 1$ has (obviously) $F = \mathbf{1}_A$. The $\sigma-$algebra generated by the X_k is the set of events that may be defined in this way. A complete proof of this would take a few pages.

1.20. Example in two dimensions: Suppose Ω is the unit square in two dimensions: $(x, y) \in \Omega$ if $0 \leq x \leq 1$ and $0 \leq y \leq 1$. The "x coordinate function" is $X(x, y) = x$. The information in this is the value of the x coordinate, but not the y coordinate. An event measurable with respect to this $\mathcal{F}$ will be any event determined by the x coordinate alone. I call such sets "bar code" sets. You can see why by drawing some. A function $f(x, y)$ is measurable with respect to the bar code algebra if there is a function $u(x)$ so that $f(x, y) = u(x)$.

1.21. Marginal density and total probability: The abstract situation is that we have a probability space, Ω with generic outcome $\omega \in \Omega$. We have some functions $(X_1(\omega), \ldots, X_n(\omega)) = X(\omega)$. With Ω in the background, we can ask for the joint distribution of $(X_1, \ldots, X_n)$. If this distribution has a probability density (ODF), we call it $u(x_1, \ldots, x_n)$. A formal definition of u would be that if $A \subseteq R^n$, then

$$P(X(\omega) \in A) = \int_{x \in A} u(x) dx \ . \tag{8}$$

Suppose we neglect the last variable, X_n, and consider the reduced vector $\tilde{X}(\omega) = (X_1, \ldots, X_{n-1})$ with probability density $\tilde{u}(x_1, \ldots, x_{n-1})$. This $\tilde{u}$ is the "marginal density" and is given by integrating u over the forgotten variable:

$$\tilde{u}(x_1, \ldots, x_{n-1}) = \int_{-\infty}^{\infty} u(x_1, \ldots, x_n) dx_n \ . \tag{9}$$

This is a continuous probability analogue of the law of total probability: integrate (or sum) over a complete set of possibilities, all values of x_n in this case.

We can prove (9) from (8) by considering a set $B \subseteq R^{n-1}$ and the corresponding set $A \subseteq R^n$ given by $A = B \times R$ (i.e. A is the set of all pairs $\tilde{x}, x_n$) with $\tilde{x} = (x_1, \ldots, x_{n-1}) \in B$). The definition of A from B is designed so that $P(X \in A) = P(\tilde{X} \in B)$. With this notation,

$$\begin{aligned} P(\tilde{X} \in B) &= P(X \in A) \\ &= \int_A u(x) dx \\ &= \int_{\tilde{x} \in B} \int_{x_n = -\infty}^{\infty} u(\tilde{x}, x_n) dx_n d\tilde{x} \\ P(\tilde{X} \in B) &= \int_B \tilde{u}(\tilde{x}) d\tilde{x} \ . \end{aligned}$$

This is exactly what it means for $\tilde{u}$ to be the PDF for $\tilde{X}$.

1.22. Classical conditional expectation: Again in the abstract setting $\omega \in \Omega$, suppose we have random variables $(X_1(\omega), \ldots, X_n(\omega))$. Now consider a function $f(x_1, \ldots, x_n)$, its expectated value $E[f(X)]$, and the conditional expectations

$$v(x_n) = E[f(X) \mid X_n = x_n] \ .$$

The Bayes' rule definition of $v(x_n)$ has some trouble because both the denominator, $P(X_n = x_n)$, and the numerator,

$$E[f(X) \cdot \mathbf{1}_{X_n = x_n}] \ ,$$

are zero.

The classical solution to this problem is to replace the exact condition $X_n = x_n$ with an approximate condition having positive (though small) probability: $x_n \leq X_n \leq x_n + \epsilon$. We use the approximaion

$$\int_{x_n}^{x_n+\epsilon} g(\tilde{x}, \xi_n) d\xi_n \approx \epsilon g(\tilde{x}, x_n) \ .$$

The error is roughly proportional to ϵ^2 and much smaller than either the terms above. With this approximation the numerator in Bayes' rule is

$$\begin{aligned} E[f(X) \cdot \mathbf{1}_{x_n \leq X_n \leq x_n+\epsilon}] &= \int_{\tilde{x} \in R^{n-1}} \int_{\xi_n = x_n}^{\xi_n = x_n+\epsilon} f(\tilde{x}, \xi_n) u(\tilde{x}, x_n) d\xi_n d\tilde{x} \\ &\approx \epsilon \int_{\tilde{x}} f(\tilde{x}, x_n) u(\tilde{x}, x_n) d\tilde{x} \ . \end{aligned}$$

Similarly, the denominator is

$$P(x_n \leq X_n \leq x_n + \epsilon) \approx \epsilon \int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x} \ .$$

If we take the Bayes' rule quotient and let $\epsilon \to 0$, we get the classical formula

$$E[f(X) \mid X_n = x_n] = \frac{\int_{\tilde{x}} f(\tilde{x}, x_n) u(\tilde{x}, x_n) d\tilde{x}}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \ . \tag{10}$$

By taking f to be the characteristic function of an event (all possible events) we get a formula for the probability density of $\tilde{X}$ given that $X_n = x_n$, namely

$$\tilde{u}(\tilde{x} \mid X_n = x_n) = \frac{u(\tilde{x}, x_n)}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \ . \tag{11}$$

This is the classical formula for conditional probability density. The integral in the denominator insures that, for each x_n, $\tilde{u}$ is a probability density as a function of $\tilde{x}$, that is

$$\int \tilde{u}(\tilde{x} \mid X_n = x_n) d\tilde{x} = 1 \ ,$$

for any value of x_n. It is very useful to notice that as a function of $\tilde{x}$, u and $\tilde{u}$ almost the same. They differ only by a constant normalization. For example, this is why conditioning Gaussian's gives Gaussians.

1.23. Modern conditional expectation: The classical conditional expectation (10) and conditional probability (11) formulas are the same as what comes from the "modern" definition from paragraph 1.6. Suppose $X = (X_1, \ldots, X_n)$ has density $u(x)$, $\mathcal{F}$ is the $\sigma-$algebra of Borel sets, and $\mathcal{G}$ is the $\sigma-$algebra generated by X_n (which might be written $X_n(X)$, thinking of X as ω in the abstract notation). For any $f(x)$, we have $\tilde{f}(x_n) = E[f \mid \mathcal{G}](x_n)$. Since $\mathcal{G}$ is generated by

X_n, the function $\tilde{f}$ being measurable with respect to $\mathcal{G}$ is the same as it's being a function of x_n. The modern definition of $\tilde{f}(x_n)$ is that it minimizes

$$\int_{R^n} \left(f(x) - \tilde{f}(x_n) \right)^2 u(x) dx \ , \tag{12}$$

over all functions that depend only on x_n (measurable in $\mathcal{G}$).

To see the formula (10) emerge, again write $x = (\tilde{x}, x_n)$, so that $f(x) = f(\tilde{x}, x_n)$, and $u(x) = u(\tilde{x}, x_n)$. The integral (12) is then

$$\int_{x_n=-\infty}^{\infty} \int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - \tilde{f}(x_n) \right)^2 u(\tilde{x}, x_n) d\tilde{x} dx_n \ .$$

In the inner integral:

$$R(x_n) = \int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - \tilde{f}(x_n) \right)^2 u(\tilde{x}, x_n) d\tilde{x} \ ,$$

$\tilde{f}(x_n)$ is just a constant. We find the value of $\tilde{f}(x_n)$ that minimizes $R(x_n)$ by minimizing the quantity

$$\int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - g \right)^2 u(\tilde{x}, x_n) d\tilde{x} = \\ \int f(\tilde{x})^2 u(\tilde{x}, x_n) d\tilde{x} + 2g \int f(\tilde{x}) u(\tilde{x}, x_n) d\tilde{x} + g^2 \int u(\tilde{x}, x_n) d\tilde{x} \ .$$

The optimal g is given by the classical formula (10).

1.24. Modern conditional probability: We already saw that the modern approach to conditional probability for $\mathcal{G} \subset \mathcal{F}$ is through conditional expectation. In its most general form, for every (or almost every) $\omega \in \Omega$, there should be a probability measure P_ω on Ω so that the mapping $\omega \to P_\omega$ is measureable with respect to $\mathcal{G}$. The measurability condition probably means that for every event $A \in \mathcal{F}$ the function $p_A(\omega) = P_\omega(A)$ is a $\mathcal{G}$ measurable function of ω. In terms of these measures, the conditional expectation $\tilde{f} = E[f \mid \mathcal{G}]$ would be $\tilde{f}(\omega) = E_\omega[f]$. Here E_ω means the expected value using the probability measure P_ω. There are many such subscripted expectations coming.

A subtle point here is that the conditional probability measures are defined on the original probability space, Ω. This forces the measures to "live" on tiny (generally measure zero) subsets of Ω. For example, if $\Omega = R^n$ and $\mathcal{G}$ is generated by x_n, then the conditional expectation value $\tilde{f}(x_n)$ is an average of f (using density u) only over the hyperplane $X_n = x_n$. Thus, the conditional probability measures P_X depend only on x_n, leading us to write P_{x_n}. Since $\tilde{f}(x_n) = \int f(x) dP_{x_n}(x)$, and $\tilde{f}(x_n)$ depends only on values of $f(\tilde{x}, x_n)$ with the last coordinate fixed, the measure dP_{x_n} is some kind of δ measure on that hyperplane. This point of view is useful in many advanced problems, but we will not need it in this course (I sincerely hope).

1.25. Semimodern conditional probability: Here is an intermediate "semimodern" version of conditional probability density. We have $\Omega = R^n$, and $\tilde{\Omega} = R^{n-1}$ with elements $\tilde{x} = (x_1, \ldots, x_{n-1})$. For each x_n, there will be a (conditional) probability density function $\tilde{u}_{x_n}$. Saying that $\tilde{u}$ depends only on x_n is the same as saying that the function $x \to \tilde{u}_{x_n}$ is measurable with respect to $\mathcal{G}$. The conditional expectation formula (10) may be written

$$E[f \mid \mathcal{G}](x_n) = \int_{R^{n-1}} f(\tilde{x}, x_n)\tilde{u}_{x_n}(\tilde{x})d\tilde{x} \ .$$

In other words, the classical $u(\tilde{x} \mid X_n = x_n)$ of (11) is the same as the semimodern $\tilde{u}_{x_n}(\tilde{x})$.

2 Gaussian Random Variables

The central limit theorem (CLT) makes Gaussian random variables important. A generalization of the CLT is Donsker's "invariance principle" that gives Brownian motion as a limit of random walk. In many ways Brownian motion is a multivariate Gaussian random variable. We review multivariate normal random variables and the corresponding linear algebra as a prelude to Brownian motion.

2.1. Gaussian random variables, scalar: The one dimensional *standard normal*, or *Gaussian*, random variable is a scalar with probability density

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

The normalization factor $\frac{1}{\sqrt{2\pi}}$ makes $\int_{-\infty}^{\infty} u(x)dx = 1$ (a famous fact). The mean value is $E[X] = 0$ (the integrand $xe^{-x^2/2}$ is antisymmetric about $x = 0$). The variance is (using integration by parts)

$$\begin{aligned}
E[X^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(x e^{-x^2/2} \right) dx \\
&= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(\frac{d}{dx} e^{-x^2/2} \right) dx \\
&= -\frac{1}{\sqrt{2\pi}} \left(x e^{-x^2/2} \right) \Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx \\
&= 0 + 1
\end{aligned}$$

Similar calculations give $E[X^4] = 3$, $E[X^6] = 15$, and so on. I will often write Z for a standard normal random variable. A one dimensional Gaussian random variable with mean $E[X] = \mu$ and variance $\text{var}(X) = E[(X - \mu)^2] = \sigma^2$ has density

$$u(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ .$$

It is often more convenient to think of Z as the random variable (like ω) and write $X = \mu + \sigma Z$. We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to express the fact that X is normal (Gaussian) with mean μ and variance σ^2. The standard normal random variable is $Z \sim \mathcal{N}(0, 1)$

2.2. Multivariate normal random variables: The $n \times n$ matrix, H, is positive definite if $x^* H x > 0$ for any n component column vector $x \neq 0$. It is symmetric if[5] $H^* = H$. A symmetric matrix is positive definite if and only if all its eigenvales are positive. Since the inverse of a symmetric matrix is symmetric, the inverse of a symmetric positive definite (SPD) matrix is also SPD. An n component random variable is a mean zero multivariate normal if it has a probability density of the form

$$u(x) = \frac{1}{z} e^{-\frac{1}{2} x^* H x} \; , \tag{13}$$

for some SPD matrix, H. We can get mean $\mu = (\mu_1, \ldots, \mu_n)^*$ either by taking $X + \mu$ where X has mean zero, or by using the density with $x^* H x$ replaced by $(x - \mu)^* H (x - \mu)$.

2.3. Characteristic functions: If $X \in R^n$ is a random variable with probability density $u(x)$, the *characteristic function* is a function of the *dual variable*, $\xi = (\xi_1, \ldots, \xi_n)$. It is defined by

$$\widehat{u}(\xi) = E\left[e^{-i\xi \cdot X}\right] = \int_{x \in R^n} e^{-i\xi \cdot x} u(x) dx \; . \tag{14}$$

We can interpret $\xi \cdot x$ as being the product of a row vector, ξ with the column vector, x, or as the dot product $\xi \cdot x = \sum_k \xi_k x_k$. The integral (14) is one form of the *Fourier transform* of u. The more abstract version $E\left[e^{-i\xi \cdot X}\right]$ is particularly useful in probability. It makes sense even when X does not have a proper probability function. The *Fourier inversion formula* (which we do not prove) states that

$$u(x) = \frac{1}{(2\pi)^n} \int_{\xi \in R^n} e^{i\xi \cdot x} \widehat{u}(\xi) d\xi \; . \tag{15}$$

The formulas (14) and (15) are similar but not the same. The characteristic function has $-i\xi \cdot x$ while the inversion formula has the opposite sign $i\xi \cdot x$. The inversion formula has a 2π factor. Different definitions of the Fourier transform put 2π factors in different places, but they must be somewhere.

2.4. Fourier transforms following Dirac: Dirac, in his book *Principles of Quantum Mechanics*, gives a simple way to remember the properties of Fourier transforms. The basic formula is,

$$\int_{-\infty}^{\infty} e^{itx} dt = 2\pi \delta(x) \; . \tag{16}$$

[5] We write H^* for the transpose of H.

For example, we can verify the Fourier inversion formula in the one dimensional case by substituting (14) into (15) and changing the order of integration:

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi x}\widehat{u}(\xi)d\xi &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i\xi x}e^{-i\xi y}u(y)dyd\xi \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{i\xi(x-y)}d\xi\right)u(y)dy \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\big(2\pi\delta(x-y)\big)u(y)dy \\
&= u(x)\ .
\end{aligned}$$

We give an informal derivation of the already informal formula (16) below.

2.5. Characteristic function of Gaussians. Characteristic functions are particularly handy for Gaussian random variables. The simplest case is the characteristic function of a standard normal, which is

$$\widehat{u}(\xi) = E\left[e^{-i\xi Z}\right] = \int_{-\infty}^{\infty} e^{-i\xi z}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\ .$$

The trick is to complete the square in the exponent:

$$-1\xi z - \frac{1}{2}z^2 = \frac{-1}{2}\left(z^2 + 2i\xi z + (i\xi)^2 - (i\xi)^2\right) = \frac{-1}{2}(z+i\xi)^2 - \xi^2/2\ .$$

Therefore,

$$\widehat{u}(\xi) = e^{-\xi^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{(z+i\xi)^2/2}dz\ .$$

It happens that this integral is independent of ξ. If we believe that, then we can take $\xi = 0$ and get

$$\widehat{u}(\xi) = e^{-\xi^2/2}\ . \tag{17}$$

We say that the characteristic function of a Gaussian is Gaussian, but this is not strictly true because there is no 2π factor on the right. Note that $1 = \widehat{u}(0)E[e^{i0\cdot Z}] = E[1]$ as it should.

If $i\xi$ were real, then the integral $\int_{-\infty}^{\infty} e^{-(z+i\xi)^2/2}dz$ would not depend on ξ because ξ would just cause a shift that does not change the area. To prove the integral is independent of ξ when $i\xi$ is not real, take the derivative with respect to ξ:

$$\frac{d}{d\xi}\int e^{-(z+i\xi)^2/2}dz = \int i(z+i\xi)e^{-(z+i\xi)^2/2}dz = i\int \frac{d}{dz}e^{-(z+i\xi)^2/2}dz = 0\ .$$

2.6. Scaling of characteristic functions: Suppose X is a univariate random variable and $Y = aX + b$. Let $\widehat{u}_X = E\left[e^{-i\xi X}\right]$ and $\widehat{u}_Y = E\left[e^{-i\xi Y}\right]$ be the corresponding characteristic functions. There is a simple formula for the characteristic function of Y in terms of that for X:

$$\begin{aligned}
\widehat{u}_Y(\xi) &= E\left[e^{-i\xi Y}\right] \\
&= E\left[e^{-i\xi(aX+b)}\right] \\
&= e^{-i\xi b} E\left[e^{-i(a\xi)X}\right] \\
\widehat{u}_Y(\xi) &= e^{-i\xi b}\widehat{u}_X(a\xi) \ .
\end{aligned}$$

In particular, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then it may be written $X = \mu + \sigma Z$, where Z is a standard normal. Therefore, the characteristic function of a $\mathcal{N}(\mu, \sigma^2)$ is

$$e^{-i\xi\mu} e^{-\sigma^2\xi^2/2} \tag{18}$$

We also could have found this formula from the integral

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-i\xi x} e^{-(x-\mu)^2/2\sigma^2} dx \ .$$

2.7. Linear transformations and characteristic functions: Calculations like these also work for multivariate random variables. If $X \in R^n$, A is an $m \times n$ matrix, and $Y \in R^m$ is given by $Y = AX$, then

$$\widehat{u}_Y(\xi) = E\left[e^{-i\xi Y}\right] = E\left[e^{-i(\xi A)\cdot X}\right] \ ,$$

so

$$\widehat{u}_Y(\xi) = \widehat{u}_X(\xi A) \ . \tag{19}$$

Note that ξ is an m component row vector and ξA is an n component row vector, as it needs to be to be the argument of $\widehat{u}_X$.

2.8. Choleski factorization: Recall the following facts from linear algebra.[6] An $n \times n$ matrix, L, is *lower triangular* if all its entries above the diagonal are zero: $L_{jk} = 0$ for $k > j$. A symmetric $n \times n$ matrix, C, is positive *semidefinite* if $x^*Cx \geq 0$ for all $x \in R^n$. The difference between positive definite and semidefinite is that a positive semidefinite matrix may be singular. If $C = LL^*$ and L is a lower triangular matrix, we say that LL^* is the *Choleski factorization* of C. A symmetric matrix is positive semidefinite if and only if it has a Choleski factorization. The L, if it exists, is unique. The matrix C is positive definite if and only if L is nonsingular. If L is nonsingular and $M = L^{-1}$, then $\left(L^*\right)^{-1} = M^{-1}$. If $C = LL^*$, then $C^{-1} = \left(L^*\right)^{-1} L^{-1} = M^*M$. Note that the transposed matrix now comes first.

In the univariate case, we get $X \sim \mathcal{N}(0, \sigma^2)$ using $X = \sigma Z$. The coefficient in the linear transformation is the square root of the coefficient, σ^2, in the

[6] See, for example, *Linear Algebra* by Gilbert Strang, or my notes *Principles of Scientific Computing*.

Gaussian probability density formula. Multivariate normal Gaussians have an $n \times n$ matrix, the *covariance matrix*, instead of the scalar σ^2. We will see that the Choleski factor, L, serves the role of σ in this case.

2.9. Standard multivariate normal: Let $Z = (Z_1, \ldots, Z_n)^*$ be a multivariate random variable whose components are independent scalar standard normals. This is the *standard multivariate normal*. Its probability density is (Here $\|x\|^2 = x^*x = \sum_k x_k^2$ and $\|\xi\|^2 = \xi\xi^* = \sum_k \xi_k^2$. The difference is because x is a column vector and ξ is a row vector.)

$$u_Z(z) = \frac{1}{(2\pi)^{n/2}} e^{-\left\|z^2\right\|^2/2} \; . \tag{20}$$

Its characteristic function is (think this through)

$$\widehat{u}_Z = e^{-\left\|\xi^2\right\|^2/2} \; . \tag{21}$$

The Fourier inversion formula (15) applied to this pair gives the identity

$$\frac{1}{(2\pi)^{n/2}} e^{-z^*z/2} = \frac{1}{(2\pi)^{n/2}} \int_{\eta \in R^n} e^{i\eta z} e^{-\eta\eta^*/2} d\eta \; . \tag{22}$$

2.10. Making multivariate normals: We have the tools to understand general multivariate normals using mappings. Suppose Z is a standard normal and $X = LZ$. The characteristic function of X is, using (19), (21), and $LL^* = C$,

$$\widehat{u}_X(\xi) = \widehat{u}_Z(\xi L) = \exp\left(\frac{-1}{2}\left(\xi L\right)^*\left(\xi L\right)\right) = e^{-\xi LL^*\xi^*/2} = e^{-\xi C\xi^*/2} \; .$$

We can use the Fourier inversion formula (15) to find the probability density $u(x)$. We will change variables in the integral from ξ to $\eta = \xi L$. This implies that $\xi = \eta L^{-1} = \eta M$ The Jacobian factor is $\det(L) d\xi = d\eta$. Using (22) with $z = Mx$, and $MM^* = C^{-1}$, the result is

$$\begin{aligned}
u(x) &= \frac{1}{(2\pi)^n} \int_{\xi \in R^n} e^{i\xi \cdot x} \exp\left(\frac{-1}{2}\left(\xi L\right)^*\right) d\xi \\
&= \frac{1}{(2\pi)^n \det(L)} \int_{eta \in R^n} e^{i\eta Mx} e^{-\eta\eta^*/2} d\eta \\
&= \frac{1}{(2\pi)^{n/2} \det(L)} e^{-x^*M^*Mx/2} \\
u(x) &= \frac{1}{(2\pi)^{n/2} \det(L)} e^{-x^*C^{-1}x/2} \qquad (23)
\end{aligned}$$

This is precisely the probability density given above for the multivariate normal (13), if $H = C^{-1}$.

To summarize, the probability density/characteristic function pair for a univariate normal is

$$u(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2\sigma^2} \iff \widehat{u}(\xi) = e^{-\sigma^2\xi^2/2} . \tag{24}$$

For a multivariate normal it is

$$u(x) = \frac{1}{\sqrt{\det(C)}} \cdot \frac{1}{(2\pi)^{n/2}} e^{-x^* C^{-1} x/2} \iff \widehat{u}(\xi) = e^{-\xi^* C\xi/2} . \tag{25}$$

The matrix C plays the role of σ^2 in both the density and the characteristic function. It is the *covariance* matrix of X, as we will see below. The normalization factor involves $\sqrt{\det C} = \det(L)$ because $\det(C) = \det(LL^*) = \det(L)\det(L^*) = \det(L)^2$.

2.11. Linear transformations of multivariate normals: If X is a multivariate normal with covariance matrix C_X and $Y = AX$, then the characteristic function of Y is given by (19):

$$\widehat{u}_Y(\xi) = \widehat{u}_X(\xi A) = \exp\left(-(\xi A) C_X (\xi A)^*\right) = e^{-\xi C_Y \xi^*} , \tag{26}$$

with

$$C_Y = A C_X A^* . \tag{27}$$

The bottom line of all this is that a linear transformation of a multivariate normal has the characteristic function of a multivariate normal and therefore is a multivariate normal. Once you know this basic fact, the covariance formula (27) is easy to derive more directly.

It is interesting if somewhat technical to consider what happens if the linear transformation A maps R^n to R^m with $m > n$. In this case, the set of values $Y = AX$ for all $X \in R^n$ is at most a hyperplane of dimension $n < M$ in R^m. That means that the distribution of Y is singular – there is no proper probability density for it. Nevertheless, Y is characterized by its perfectly proper characteristic function (26).

2.12. Moment generating functions: The characteristic function of a random variable is closely related to the *moment generating function*, or simply *generating function* $E\left[e^{\lambda X}\right]$. Just take $\lambda = -i\xi$. As with the moment generating function, the characteristic function determines the moments of X. For example,

$$\partial_x \widehat{u}_X(\xi) = \partial_\xi E\left[e^{-i\xi X}\right] = E\left[\partial_\xi e^{-i\xi X}\right] = -iE\left[X e^{-i\xi X}\right] .$$

If ξ has more than one component, this is a vector equation. In particular, setting $\xi = 0$ gives

$$i\partial_x \widehat{u}_X(\xi)\Big|_{\xi=0} = -E[X] . \tag{28}$$

We can apply this to get second moments as well. The result is (Please check the signs, which are detemined by several factors of $-i$.)

$$E\left[X_j X_k\right] = -\partial_{\xi_j}\partial_{\xi_k}\widehat{u}_X(\xi)\Big|_{\xi=0} \; . \tag{29}$$

In particular, for the Gaussian generating function (25), (29) gives

$$E[X_j X_k] - C_{jk} \; . \tag{30}$$

This says that the entries of C are the covariances of the components of the multivariate normal, X.

2.13. Independence and decorrelation: Suppose X_1 and X_2 are random variables with a joint PDF $u(x_1, x_2)$. Suppose they both have mean zero, though this is only for convenience. The covariance is $\mathrm{cov}(X_1, X_2) = E[X_1 X_2]$. If the covariance is zero, we say X_1 and X_2 are *uncorrelated*. If $u(x_1, x_2) = u_1(x_1)u_2(x_2)$, then X_1 and X_2 are independent. It is easy to see that independent random variables are uncorrelated but uncorrelated variables need not be independent.[7]

The exception is Gaussian random variables. Let X be a multivariate normal with mean zero. If the components of X are uncorrelated then $C_{jk} = 0$ for $j \neq k$. Therefore C and $H = C^{-1}$ are diagonal and the expression x^*Hx in (25) may be written $\sum_k h_{kk}x_k^2$. This implies that

$$u(x) = Const \cdot e^{x^*Hx} = Const \cdot \prod_k e^{-h_{kk}x_k^2/2} \; ,$$

which in turn implies that the X_k are independent.

2.14. Principal component analysis: Suppose the eigenvalues and eigenvectors of C are $Cv_j = \lambda_j v_j$. The v_j are the *principal vectors*, or *principle directions*, and the λ_j are the *principle values* of C. The eigenvalue problem for C is *principal component analysis* or *PCA*. We can express $x \in R^n$ as a linear combination of the v_j,

$$x = \sum_{j=1}^{n} y_j v_j \; . \tag{31}$$

This is the principal compenent expansion of x. The weights y_j are the *principal components* of x with respect to the covariance matrix C. The relations (31) may be expressed in matrix form as $x = Vy$, where V is the $n \times n$ matrix whose columns are the v_j and $y = (y_1, \ldots, y_n)^*$. Since the eigenvectors of a symmetric matrix are orthogonal to each other, we may normalize them so that $v_j^* v_k = \delta_{jk}$, which is the same as saying that V is an orthogonal matrix, $V^*V = I$. This leads to formula for the principal components in (31):

$$y_j = v_j^* x \; . \tag{32}$$

[7]For example, if X_1, X_2 is uniformly distributed in the unit circle, then X_1 and X_2 are uncorrelated, but if $|X_1| > .9$ then $|X_2| < .5$ (draw a picture), so they are not independent.

In matrix form, since $V^* = V^{-1}$, we have $y = V^*x$. The matrix form of the eigenvalue eigenvector relations is $CV = V\Lambda$, which may be rewritten $V^*CV = \Lambda$.

2.15. PCA variables: Let X be a multivariate normal with covariance matric C and $Y = V^*X$. The components Y_j are the principal components of X. We will see that they are independent normals with variances $\sigma_j^2 = \lambda_j$. First, Y is a multivariate normal, since it is a linear transformation of X. Then we use (27), with $A = V^*$ to calculate

$$C_Y = V^*CV = \Lambda \; .$$

This implies that the components of Y are uncorrelated and therefore independent.

2.16. Central Limit Theorem: Let X be an n dimensional random variable with probability density $u(x)$. Let $X^{(1)}$, $X^{(2)}$, ..., be a sequence of independent samples of X, that is, independent random variables with the same density u. Statisticians call this iid (independent, identically distributed). If we need to talk about the individual components of $X^{(k)}$, we write $X_j^{(k)}$ for component j of $X^{(k)}$. For example, suppose we have a population of people. If we choose a person "at random" and record his or her height (X_1) and weight (X_2), we get a two dimensional random variable. If we measure 100 people, we get 100 samples, $X^{(1)}$, ..., $X^{(100)}$, each consisting of a height and weight pair. The weight of person 27 is $X_2^{(27)}$. Let $\mu = E[X]$ be the mean and $C = E[(X-\mu)(X-\mu)^*]$ the covariance matrix. The Central Limit Theorem (CLT) states that for large n, the random variable

$$R^{(n)} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (X^{(k)} - \mu)$$

has a probability distribution close to the multivariate normal with mean zero and covariance C. One interesting consequence is that if X_1 and X_2 are uncorrelated then an average of many independent samples will have $R_1^(n)$ and $R_2^{(n)}$ nearly independent.

2.17. What the CLT says about Gaussians: The Central Limit Theorem tells us that if we avarage a large number of independent samples from the same distribution, the distribution of the average depends only on the mean and covariance of the starting distribution. It may be surprising that many of the properties that we deduced from the formula (**??**) may be found with almost no algebra simply knowing that the multivariate normal is the limit of averages. For example, we showed (or didn't show) that if X is multivariate normal and $Y = AX$ where the rows of A are linearly independent, then Y is multivariate normal. This is a consequence of the averaging property. If X is (approximately) the average of iid random variables U_k, then Y is the average of random variables $V_k = AU_k$. Applying the CLT to the averaging of the V_k shows taht Y is also multivariate normal.

Now suppose U is a univariate random variable with iid samples U_k, and $E[U_k] = 0$, $E[U_k^2 = \sigma^2]$, and $E[U_k^4] = a_4 < \infty$ Define $X_n = \frac{1}{\sqrt{n}} \sum_{k=n}^{n} U_k$. A calculation shows that $E[X_n^4] = 3\sigma^4 + \frac{1}{n} a_4$. For large n, the fourth moment of the average depends only on the second moment of the underlying distribution. A multivariate and slightly more general version of this calculation gives "Wick's theorem", an expression for the expected value of a product of components of a multivariate normal in terms of covariances.