PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

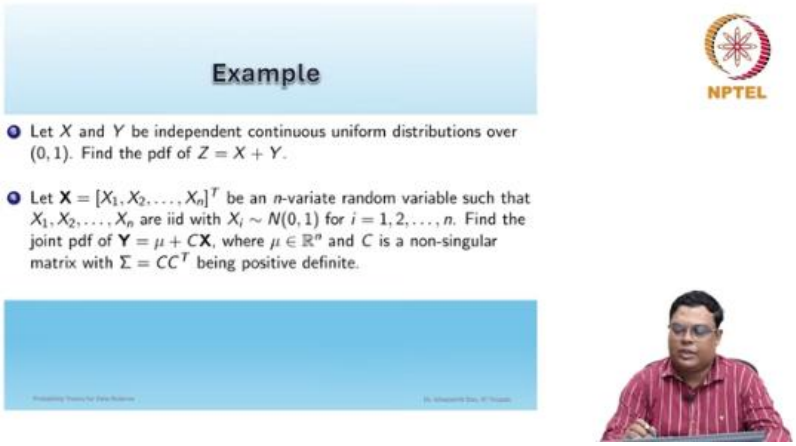
Indian Institute of Technology Tirupati

Week - 11

Lecture - 57

Convolution and Example on Transformation of n-variate Random Variables

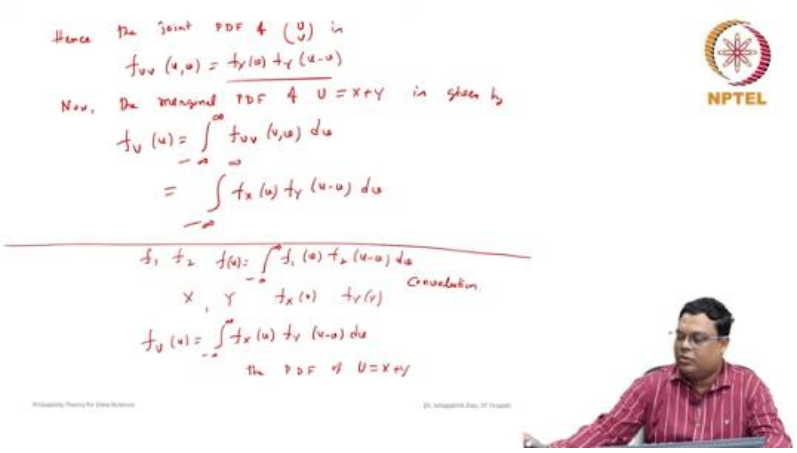
Let us discuss this example. Let X and Y be independent continuous uniform distributions over the interval from 0 to 1. Find the probability density function of J, where J = X + Y. So, in this theory, note that if you have two functions, f_1 and f_2, and you perform an integration where f_1 is multiplied by f_2 shifted by a variable u, and then integrated over the entire range from $-\infty$ to ∞, this will result in a function of u. This operation is known as the convolution of the two functions.



So, here we have two density functions, x and y, and we are performing the convolution. However, there is a particular significance in this process. By this convolution, you have two random variables, x and y, with their respective probability density functions, f(x) and f(y). If you perform the convolution between these two functions, the result, $f_u(u)$, is equal

to the integral from -∞ to ∞ of $f_x(v)$ multiplied by $f_\gamma(u - v)$ with respect to v. So, this convolution gives the probability density function of U, where U = X + Y.

That is why this convolution has a particular significance in the transformation of random variables in probability theory. By using convolution, we can find the probability density function of the sum of two independent continuous random variables. Whenever there are two independent random variables, we can apply the convolution to determine their sum's probability density function. So, in this example, in this particular case, let X and Y be two independent random variables. So, X and Y are two independent uniform random variables, each uniformly distributed over the interval [0, 1].



Since they are independent, the probability density functions of X and Y are given by: $f_x(x) = 1$ when $x \in [0, 1]$, and 0 otherwise. Similarly, $f_\gamma(y) = 1$ when $y \in [0, 1]$, and 0 otherwise. Now, using the previous theorem, the probability density function of U, where U = X + Y, can be determined. In this question, it is given that X and Y are independent continuous uniform distributions over the interval [0, 1].

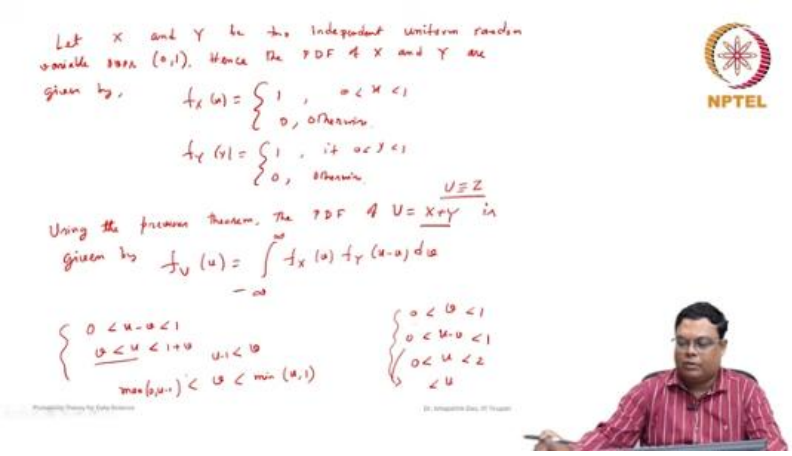
Find the probability density function of Z, where Z = X + Y. In the theory, we used the notation U, so let's use the same notation to avoid confusion. Although the question uses the notation Z, we will use U for consistency. The probability density function of U, which is equal to X + Y, can be found by integrating from -∞ to ∞. Note that we could have used a different transformation, such as letting v = x, but ultimately we would reach the same result.

This theory helps simplify the formula. The probability density function is obtained by multiplying the probability density functions of X and Y, with one of the functions shifted, and then integrating with respect to v. This value will be 1 only if $v \in [0, 1]$, because that is the condition here. Additionally, $u - v \in [0, 1]$, meaning $u - v$ must lie within this range. Now, let's think about the value of u.

Since $u = x + y$, the minimum value of u will be 0, because both x and $y \in [0, 1]$. The maximum value of u will be 2. So, these three conditions must be satisfied for the value to be non-zero; otherwise, it will be zero. Now, simplifying the situation, we can see that the lower bound for v is 0. Looking at this relationship, if we want to integrate with respect to v, $u - v \in [0, 1]$.

By adding v to both sides of the inequality, we can conclude that $u \geq 1 + v$. From this relationship, we can conclude that $v < u$, and also $v > u - 1$. So, v must satisfy both of these conditions. Therefore, v has to be greater than max(0, u - 1). Additionally, v must be less than min(u, 1).

So, the condition is that v must be greater than max(0, u - 1), and less than min(u, 1). This condition will change depending on the values of u. As you can see, the minimum and maximum values of v will vary accordingly. Let's continue on the next page. We found that $v \in [0, 2]$, and v must be at least 0 and less than u - 1.

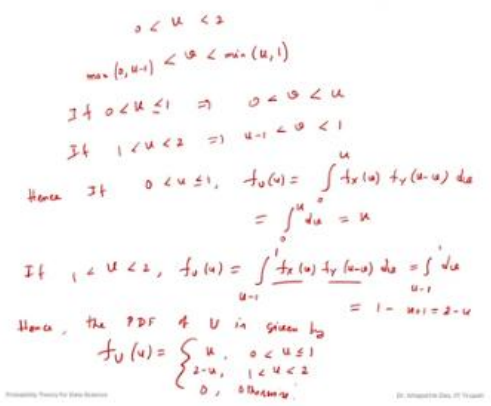


The maximum value of v is the min(u - 1, 1), which must satisfy both conditions. So, depending on the values of u, this limit will change. For example, if u ∈ [0, 1], v must be greater than 0 and less than min(u, 1). Since u - 1 is negative or 0 in this case, the maximum value of v will be 0. Therefore, v ∈ (0, 0).

If u ≤ 1, the min(u, 1) is just u. If u ∈ (1, 2), then v's limit will be u - 1, since max(0, u - 1) = u - 1. So, u - 1 < min(u, 1) because, when u > 1, the min will be 1. This is the integration we need to perform. Finally, what are we getting?

We need to carry out this integration, and depending on the different values of u, the limit of v will change. In this range, both values are 1, and this is from 0 to u, which results in u. Now, if u ∈ (1, 2), the probability density function for u is the integral from u - 1 to 1, and then the product of the probability density function for x and the probability density function for y, evaluated at u - v, with respect to v. In this range, this value will also be 1. So, this is $\int_{xy}(u - 1)^1 f_x(v) * f_y(u - v)\, dv$. If you simplify this, it becomes 1 - u + 1, which is 2 - u. Otherwise, the value is 0. Hence, we can write that the probability density function of u is as follows:

It is equal to u if u ∈ (0, 1], equal to 2 - u if u ∈ (1, 2), and equal to 0 otherwise. If you were to graph this function, it would look like this. So, this is 0, this is 1, and this is 2. Suppose this is the value we're considering. You can see how this density function looks: it is 0 initially, but from 0 to 1, the function f(u) = u. Up to 1, the value increases.



At 1, the value is 1. From 1 to 2, the function is 2 - u. At 2, the function reaches 0. It decreases linearly until it reaches 0 at 2. So, this is the probability density function for u. Otherwise, the function is 0.

This is the joint probability density function, which is the density function for u, where u = x + y, as was asked. So, this is the density function. I hope you understand that this is one example of convolution. To use this theorem, it is given that if x and y are two independent continuous random variables with probability density functions f_X(x) and f_Y(y), respectively, then the probability density function of u, where u = x + y, is given by this relationship. We have used this to find the solution.

There are many other examples as well. You can refer to some textbooks and practice them. Initially, it may be a little complicated to determine the range and perform the integration. If you make a mistake in determining the range, you will not get the correct answer. That is why you need to practice more examples to understand the process and perform it easily.

So, practice is key. Now, we will discuss another example. In most cases, we have used bivariate transformations. We will discuss one example here because considering more than two variables would be very complicated. In this example, we will discuss a multivariate transformation. Let us consider the following: $X = (X_1, X_2, ..., X_n)$ is a random variable such that $X_1, X_2, ..., X_n$ are independent and identically distributed (i.i.d.), with each X_i being normal with a mean of 0 and variance of 1, for i = 1 to n.

We need to find the joint probability density function of $Y = \mu + C * X$, where μ is a real number, C is a non-singular matrix, and $\Sigma = C * C^T$ is a positive definite matrix. Now, let us proceed. It is given that $X = (X_1, X_2, ..., X_n)$ is a multivariate random variable, where $X_1, X_2, ..., X_n$ are i.i.d.

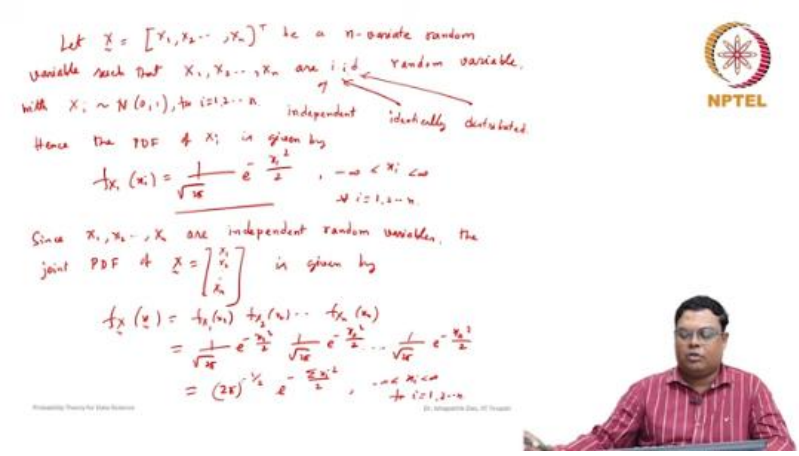
But what does i.i.d. mean? It means that $X_1, X_2, ..., X_n$ are independent and identically distributed. So, 'i.i.d.' stands for independent and identically distributed. The 'i' stands for independent, the second 'i' stands for identically, and the 'd' stands for distributed.

In other words, the random variables are independent, meaning they do not affect each other, and they are identically distributed, meaning they all follow the same probability distribution. In this example, we are dealing with random variables that follow the standard normal distribution. Each variable, X_i, is normally distributed with a mean of 0 and a variance of 1.

Specifically, each X_i follows the standard normal distribution for i ranging from 1 to n. Since the variables are independent, the probability density function (PDF) of X_i is given by:
$f\_X_i(x_i) = (1 / \sqrt{(2\pi)}) * e^{(-x_i^2 / 2)}$, where $x_i \in (-\infty, \infty)$.
Since the random variables are independent, for all values of i from 1 to n, the joint probability density function of X_1, X_2, ..., X_n is given by the product of their individual probability density functions.

So, the joint probability density function will be the product of the probability density function of X_1, X_2, ..., X_n. This means that each probability density function will follow the form:
$(1 / \sqrt{(2\pi)}) * e^{(-x_1^2 / 2)}$,
$(1 / \sqrt{(2\pi)}) * e^{(-x_2^2 / 2)}$,
...,
$(1 / \sqrt{(2\pi)}) * e^{(-x_n^2 / 2)}$.
When we combine all of these together, we get a product involving $(2\pi)^{(-n/2)}$ and multiplied by $e^{(-\Sigma x_i^2 / 2)}$, where each $x_i \in (-\infty, \infty)$.

This is the joint probability density function. Since we need to perform a transformation of random variables, this is the joint probability density function we have derived. In this question, they have asked us to consider the transformation where $Y = \mu + C * X$. Let's take this transformation where $Y = G * X$. We will use the tilde symbol because Y and X represent vectors.



If we do not use the tilde symbol, we have already accounted for it. Since y is a vector, the transformation G * X = μ + C * X, where μ is a real number and is an element in $\mathbb{R}^n$ (a real number space). C is a non-singular matrix. Since we are performing a transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, the matrix C will be an n × n matrix. If we take the transpose of C, we get C^T, and the resulting matrix would be used for further operations.

So, C is a non-singular matrix, and μ is just an element in $\mathbb{R}^n$, not a random variable. This is a vector transformation, so we can refer to it as a transformation applied to the vector X. The term μ is a matrix with dimensions n × 1, and this results in a vector with dimensions n × 1. The matrix C and its transpose C^T are both square matrices of size n × n, so the result will also be a square matrix of the same size. Now, the matrix Σ is a positive definite matrix. I assume you are familiar with the concept of a positive definite matrix.

To explain, suppose matrix A is a square matrix of size n × n. It is called positive definite if, when you take any vector x and multiply it by the matrix A, the result is always greater than or equal to 0. This holds true for all vectors x belonging to $\mathbb{R}^n$. In this case, the vector x has dimensions n × 1, and the matrix A is of size n × n, which results in a scalar value. This scalar value will always be greater than or equal to 0, but not every matrix satisfies this condition.

If a matrix satisfies this relationship, it is considered a positive definite matrix. To verify that a matrix is positive definite, we use equivalent conditions. For example, if a matrix is positive definite, all of its eigenvalues must be non-negative. Specifically, all the

eigenvalues must be positive. It is important to note that a positive definite matrix will always result in a value strictly greater than 0 when applied to any non-zero vector.

It will only equal 0 when the vector is exactly 0. So, an equivalent way of saying that a matrix is positive definite is that all of its eigenvalues are positive. For the multivariate normal distribution, we also use the concept of a positive definite matrix. However, we assume you are already familiar with the concept of a positive definite matrix, so we won't go into the details right now. If necessary, we can discuss it again later.

In this case, Σ is a positive definite matrix, which is required for the multivariate normal distribution. This is part of a transformation that ultimately leads us to the multivariate normal distribution density function, and we will explore how this works. Now, let's discuss the inverse. If we have the equation where $y = g(X)$, which is a transformation involving μ and $C * X$, where X belongs to $\mathbb{R}^n$, we can see that g is a transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$. Since C is a non-singular matrix, its inverse exists.

From the equation, we can rearrange to express X as $C^{-1} * (y - \mu)$. This is the inverse transformation, denoted as h(y). So, the inverse exists here, and now we need to determine the Jacobian. To understand how to find the Jacobian, let's consider a simple example. For simplicity, we will look at the case of transforming from $\mathbb{R}^2$ to $\mathbb{R}^2$, and we will start with a basic example of transforming from $\mathbb{R}^2$ to $\mathbb{R}^2$.

Suppose we have a matrix with elements A_{11}, A_{12}, A_{21}, and A_{22}. Now, consider a transformation where $y_1 = \mu_1$, and $y_2 = \mu_2$. The relationship for this transformation is as follows:

$x_1 = A_{11} * (y_1 - \mu_1) + A_{12} * (y_2 - \mu_2)$

$x_2 = A_{21} * (y_1 - \mu_1) + A_{22} * (y_2 - \mu_2)$.

Next, we want to find the Jacobian for this transformation.

The Jacobian is defined by the partial derivatives of x_1 and x_2 with respect to y_1 and y_2. To calculate this, we can look at the following partial derivatives:

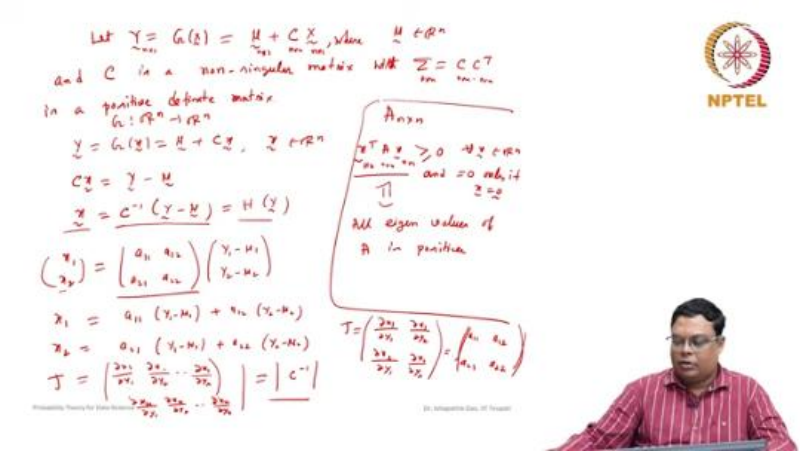
The derivative of x_1 with respect to y_1 will simply be A_{11}, since μ_1 is a constant. Similarly, the derivative of x_1 with respect to y_2 will be A_{12}. The derivative of x_2 with respect to y_1 will be A_{21}, and the derivative of x_2 with respect to y_2 will be A_{22}.

Please check this yourself for clarity. If you perform the integration and differentiation, you will obtain this result. The derivative of x_2 with respect to y_2 will give you A_{22}. Therefore, the Jacobian matrix ends up being the same matrix as the one we started with. You can observe this.
Next, if you take the determinant of the Jacobian matrix, it will yield the determinant of this same matrix.

We denote the Jacobian by J, which represents the determinant of the matrix. Since this is a transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, the concept can be extended, though notationally it might become a bit more complicated. However, the result remains the same. Thus, if you calculate the Jacobian for this case, J will be the determinant of the matrix.

You will need to write out the partial derivatives, such as the derivative of X_1 with respect to Y_1, the derivative of X_2 with respect to Y_1, and so on, up to the derivatives involving X_n and Y_n.

Ultimately, you will find that the Jacobian is simply the determinant of C^{-1}. Hopefully, you understand this. Otherwise, to extend the concept from $\mathbb{R}^n$ to $\mathbb{R}^n$, you just need to add more terms to the process, with the final result being the determinant of C^{-1}. Now, since $\Sigma = C * C^T$, we know that the determinant of C^{-1} is not just 1 divided by C. The determinant of C is 1 over the determinant of C. Also, since $\Sigma = C * C^T$, we can calculate the determinant of Σ.



The determinant of Σ will be the determinant of C * C^T. Now, for square matrix multiplication, we know that the determinant of a product is the product of the determinants. Additionally, the determinant of a transpose is equal to the determinant of the original matrix. Thus, the determinant of C * C^T is the same as the determinant of C^2. Therefore, the determinant of C will be √(det(Σ)).

Finally, we can express this as 1 / √(det(Σ)). So, the inverse of C can be written as det(Σ)^(-1/2). This computation is necessary, and I hope you were able to follow it. Now, let's move on. Since Σ is a positive definite matrix, all its eigenvalues are positive. Therefore, the determinant of Σ will always be positive, and the square root of the determinant will also be positive. As a result, this Jacobian is always positive and greater than zero for any value of y.

This concludes our discussion on the Jacobian and the transformation. Now, let's focus on the joint probability density function of y. We began with the transformation where y = μ + C * X, and we want to find the joint probability density function for y. According to the theorem, the probability density function of y is the absolute value of the Jacobian multiplied by the probability density function of X, evaluated at the inverse of y.

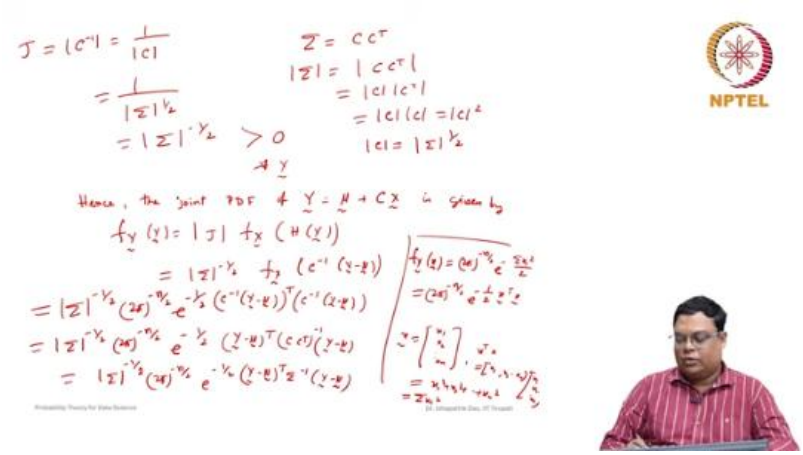
From what we have already worked through, we know that the Jacobian of the transformation involves Σ^(-1/2). Now, to determine f(X), we refer back to the function we wrote earlier. If we substitute y, we get that X = C^{-1} * (y - μ). This means we need to write the probability density function of X evaluated at the inverse transformation of y, which is

$C^{-1} * (y - \mu)$. We know that the probability density function of X has a specific form, which includes a constant factor involving 2π raised to the power of -n/2, along with an exponential function of the sum of the squares of the components of X.

This is the general form of the probability density function for X. When we treat X as a vector, the sum of the squares of its components (like X_1, X_2, ..., Xn) can be expressed as the square of the vector. So, when we substitute X with $C^{-1} * (y - \mu)$, we get a similar expression where the sum of the squares of the components of $C^{-1} * (y - \mu)$ is used. Finally, this leads us to the joint probability density function of y in the transformed form, which involves this new representation. Finally, this leads us to the joint probability density function of y in the transformed form, which involves this new representation.

This involves a factor related to the covariance matrix, raised to the power of -1/2, and multiplied by a factor that depends on π raised to the power of -n/2. Additionally, there is an exponential factor that includes a negative half, which is related to the distance between the transformed vector and the mean vector, expressed in terms of the covariance matrix. Now, we express this in terms of the difference between the transformed vector and the mean, taking the transpose of this difference. This is then multiplied by the inverse of the matrix C. The result is the inverse of the product of C and its transpose, which represents the covariance matrix Σ.

Finally, we end up with an expression involving the determinant of the covariance matrix raised to the power of -1/2, a factor involving π raised to the power of -n/2, and an exponential function involving the distance between the transformed vector and the mean, scaled by the inverse of the covariance matrix. This process gives us the probability density function for a multivariate distribution, valid for any vector y in $\mathbb{R}^n$. This is a well-known form for the density function of a multivariate normal distribution. You may recall that earlier we computed something related to this. While it would take time to revisit that point, you can refer to the previous video where we discussed it.



Essentially, this is the joint density function of a multivariate normal distribution with mean μ and covariance matrix Σ. Therefore, the transformation where Y = μ + C * X follows a multivariate normal distribution. The expected value of Y is E(Y) = μ, and the covariance of Y is Cov(Y) = Σ. The relationship between Σ and C is that $\Sigma = C * C^T$. This is the result we obtained by using the transformation of the random variable.