MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J Fall 2018
Lecture 15

MULTIVARIATE NORMAL DISTRIBUTIONS (CTD.); CHARACTERISTIC FUNCTIONS

Contents

1 EQUIVALENCE OF THE THREE DEFINITIONS OF THE MULTIVARIATE NORMAL DISTRIBUTION

1.1 The definitions

Recall the following three definitions from the previous lecture.

Definition 1. *A random vector* $\mathbf{X}$ *has a* **nondegenerate (multivariate) normal** *distribution if it has a joint PDF of the form*

$$f_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\Big\{ -\frac{(\mathbf{x}-\mu)V^{-1}(\mathbf{x}-\mu)^T}{2}\Big\},$$

for some real vector μ *and some positive definite matrix* V.

Definition 2. *A random vector* $\mathbf{X}$ *has a* **(multivariate) normal** *distribution if it can be expressed in the form*

$$\mathbf{X} = D\mathbf{W} + \mu,$$

for some matrix D *and some real vector* μ*, where* $\mathbf{W}$ *is a random vector whose components are independent* $N(0,1)$ *random variables.*

Definition 3. *A random vector* $\mathbf{X}$ *has a* **(multivariate) normal** *distribution if for every real vector* $\mathbf{a}$*, the random variable* $\mathbf{a}^T\mathbf{X}$ *is normal.*

2 PROOF OF EQUIVALENCE

In the course of the proof of Theorem 1 in the previous lecture, we argued that if $\mathbf{X}$ is multivariate normal, in the sense of Definition 2, then:

(a) It also satisfies Definition 3: if $\mathbf{X} = D\mathbf{W} + \mu$, where the W_i are independent, then $\mathbf{a}^T\mathbf{X}$ is a linear function of independent normals, hence normal.

(b) As long as the matrix D is nonsingular (equivalently, if $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = DD^T$ is nonsingular), $\mathbf{X}$ also satisfies Definition 1. (We used the derived distributions formula.)

We complete the proof of equivalence by establishing converses of the above two statements.

Theorem 1.

(a) If $\mathbf{X}$ *satisfies Definition 1, then it also satisfies Definition 2.*

(b) If $\mathbf{X}$ *satisfies Definition 3, then it also satisfies Definition 2.*

Proof:

(a) Suppose that $\mathbf{X}$ satisfies Definition 1, so in particular, the matrix V is positive definite. Let D be a symmetric matrix such that $D^2 = V$. Since

$$(\det(D))^2 = \det(D^2) = \det(V) > 0,$$

we see that D is nonsingular, and therefore invertible. Let

$$\mathbf{W} = D^{-1}(\mathbf{X} - \mu).$$

Note that $\mathbf{E}[\mathbf{W}] = 0$. Furthermore,

$$\begin{aligned}
\text{Cov}(\mathbf{W}, \mathbf{W}) &= \mathbf{E}[D^{-1}(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T D^{-1}] \\
&= D^{-1}\mathbf{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]D^{-1} \\
&= D^{-1}VD^{-1} = I.
\end{aligned}$$

We have shown thus far that the W_i are normal and uncorrelated. We now proceed to show that they are independent. Using the formula for the PDF of $\mathbf{X}$ and the change of variables formula, we find that the PDF of $\mathbf{W}$ is of the form

$$c \cdot \exp\{-\mathbf{w}^T\mathbf{w}/2\} = c \cdot \exp\{(w_1^2 + \cdots + w_n)^2/2\},$$

for some normalizing constant c, which is the joint PDF of a vector of independent normal random variables. It follows that $\mathbf{X} = D\mathbf{W} + \mu$ is a multivariate normal in the sense of Definition 2.

(b) Suppose that $\mathbf{X}$ satisfies Definition 3, i.e., any linear function $\mathbf{a}^T\mathbf{X}$ is normal. Let $V = \text{Cov}(\mathbf{X}, \mathbf{X})$, and let D be a symmetric matrix such that $D^2 = V$. We first give the proof for the easier case where V (and therefore D) is invertible.

Let $\mathbf{W} = D^{-1}(\mathbf{X} - \mu)$. As before, $\mathbf{E}[\mathbf{W}] = 0$, and $\text{Cov}(\mathbf{W}, \mathbf{W}) = I$. Fix a vector $\mathbf{s}$, Then, $\mathbf{s}^T\mathbf{W}$ is a linear function of $\mathbf{W}$, and is therefore normal. Note that

$$\text{var}(\mathbf{s}^T\mathbf{W}) = \mathbf{E}[\mathbf{s}^T\mathbf{W}\mathbf{W}^T\mathbf{s}] = \mathbf{s}^T\text{Cov}(\mathbf{W}, \mathbf{W})\mathbf{s} = \mathbf{s}^T\mathbf{s}.$$

Since $\mathbf{s}^T\mathbf{W}$ is a scalar, zero mean, normal random variable, we know that

$$M_{\mathbf{W}}(\mathbf{s}) = \mathbf{E}[\exp\{\mathbf{s}^T\mathbf{W}\}] = M_{\mathbf{s}^T\mathbf{W}}(1) = \exp\{\text{var}(\mathbf{s}^T\mathbf{W})/2\} = \exp\{\mathbf{s}^T\mathbf{s}/2\}.$$

We recognize that this is the transform associated with a vector of independent standard normal random variables. By the inversion property of transforms, it follows that $\mathbf{W}$ is a vector of independent standard normal random variables. Therefore, $\mathbf{X} = D\mathbf{W} + \mu$ is multivariate normal in the sense of Definition 2.

(b)′ Suppose now that V is singular (as opposed to positive definite). For simplicity, we will assume that the mean of $\mathbf{X}$ is zero. Then, there exists some $\mathbf{a} \neq 0$, such that $V\mathbf{a} = 0$, and $\mathbf{a}^T V\mathbf{a} = 0$. Note that

$$\mathbf{a}^T V\mathbf{a} = \mathbf{E}\big[(\mathbf{a}^T\mathbf{X})^2\big].$$

This implies that $\mathbf{a}^T\mathbf{X} = 0$, with probability 1. Consequently, some component of $\mathbf{X}$ is a deterministic linear function of the remaining components.

By possibly rearranging the components of $\mathbf{X}$, let us assume that X_n is a linear function of $(X_1, \ldots, X_{n-1})$. If the covariance matrix of $(X_1, \ldots, X_{n-1})$ is also singular, we repeat the same argument, until eventually a nonsingular covariance matrix is obtained. At that point we have reach the situation where $\mathbf{X}$ is partitioned as $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, with $\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) > 0$, and with $\mathbf{Z}$ a linear function of $\mathbf{Y}$ (i.e., $\mathbf{Z} = A\mathbf{Y}$, for some matrix A, with probability 1).

The vector $\mathbf{Y}$ also satisfies Definition 3. Since its covariance matrix is nonsingular, the previous part of the proof shows that it also satisfies Definition 2. Let k be the dimension of $\mathbf{Y}$. Then, $\mathbf{Y} = D\mathbf{W}$, where $\mathbf{W}$ consists of k independent standard normals, and D is a $k \times k$ matrix. Let $\overline{\mathbf{W}}$ be a vector of $n - k$ independent standard normals. Then, we can write

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} D & 0 \\ AD & 0 \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \overline{\mathbf{W}} \end{bmatrix},$$

which shows that $\mathbf{X}$ satisfies Definition 2.

We should also consider the extreme possibility that in the process of eliminating components of $\mathbf{X}$, a nonsingular covariance matrix is never obtained. But in that case, we have $\mathbf{X} = 0$, which also satisfies Definition 2, with $D = 0$. (This is the most degenerate case of a multivariate normal.)

□

3 WHITENING OF A SEQUENCE OF NORMAL RANDOM VARIABLES

The last part of the proof in the previous section provides some interesting intuition. Given a multivariate normal vector $\mathbf{X}$, we can always perform a change of coordinates, and obtain a representation of that vector in terms of independent normal random variables. Our process of going from $\mathbf{X}$ to $\mathbf{W}$ involved factoring the covariance matrix V of $\mathbf{X}$ in the form $V = D^2$, where D was a symmetric square root of V. However, other factorizations are also possible. The most useful one is described below.

Let

$$\begin{aligned} W_1 &= X_1, \\ W_2 &= X_2 - \mathbf{E}[X_2 \mid X_1], \\ W_3 &= X_3 - \mathbf{E}[X_3 \mid X_1, X_2], \\ \vdots & \quad \vdots \\ W_n &= X_n - \mathbf{E}[X_n \mid X_1, \ldots, X_{n-1}]. \end{aligned}$$

(a) Each W_i can be interpreted as the new information provided by X_i, given the past, $(X_1, \ldots, X_{i-1})$. The W_i are sometimes called the **innovations**.

(b) When we deal with multivariate normals, conditional expectations are linear functions of the conditioning variables. Thus, the W_i are linear functions of the X_i. Furthermore, we have $\mathbf{W} = L\mathbf{X}$, where L is a lower triangular matrix (all entries above the diagonal are zero). The diagonal entries of L are all equal to 1, so L is invertible. The inverse of L is also lower triangular. This means that the transformation from $\mathbf{X}$ to $\mathbf{W}$ is **causal** (W_i can be determined from $X_1, \ldots, X_i$) and **causally invertible** (X_i can be determined from $W_1, \ldots, W_i$). Engineers sometimes call this a "causal and causally invertible whitening filter."

(c) The W_i are independent of each other. This is a consequence of the general fact $\mathbf{E}[(X - \mathbf{E}[X \mid Y])Y] = 0$, which shows that the W_i is uncorrelated with $X_1, \ldots, X_{i-1}$, hence uncorrelated with $W_1, \ldots, W_{i-1}$. For multivariate normals, we know that zero correlation implies independence. As long as the W_i have nonzero variance, we can also normalize them so that their variance is equal to 1.

(d) The covariance matrix of $\mathbf{W}$, call it B, is diagonal. An easy calculation shows that $\mathrm{Cov}(X, X) = L^{-1}B(L^{-1})^T$. This kind of factorization into a product of a lower triangular ($L^{-1}B^{1/2}$) and upper triangular ($B^{1/2}(L^{-1})^T$) matrix is called a **Cholesky factorization.**

4 INTRODUCTION TO CHARACTERISTIC FUNCTIONS

We have defined the moment generating function $M_X(s)$, for real values of s, and noted that it may be infinite for some values of s. In particular, if $M_X(s) = \infty$ for every $s \neq 0$, then the moment generating function does not provide enough information to determine the distribution of X. (As an example, consider a PDF of the form $f_X(x) = c/(1 + x^2)$, where c is a suitable normalizing constant.) A way out of this difficulty is to consider **complex values** of s, and in particular, the case where s is a purely imaginary number: $s = it$, where $i = \sqrt{-1}$, and $t \in \mathbb{R}$. The resulting function is called the **characteristic function**, formally defined by

$$\phi_X(t) = \mathbf{E}[e^{itX}].$$

For example, when X is a continuous random variable with PDF f, we have

$$\phi_X(t) = \int e^{ixt} f(x)\, dx,$$

which very similar to the Fourier transform of f (except for the absence of a minus sign in the exponent). Thus, the relation between moment generating functions and characteristic functions is of the same kind as the relation between Laplace and Fourier trasnforms.

Note that e^{itX} is a **complex-valued** random variable, a new concept for us. However, using the relation $e^{itX} = \cos(tX) + i\sin(tX)$, defining its expectation is straightforward:

$$\phi_X(t) = \mathbf{E}[\cos(tX)] + i\mathbf{E}[\sin(tX)].$$

We make a few key observations:

(a) Because $|e^{itX}| \le 1$ for every t, its expectation, $\phi_X(t)$ is well-defined and finite for every $t \in \mathbb{R}$. In fact, $|\phi_X(t)| \le 1$, for every t.

(b) The key properties of moment generating functions (cf. Lecture 14) are also valid for characteristic functions (same proof).

Theorem 2.

(a) If $Y = aX + b$, then $\phi_Y(t) = e^{itb}\phi_X(at)$.

(b) If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

(c) Let X and Y be independent random variables. Let Z be equal to X, with probability p, and equal to Y, with probability $1 - p$. Then,

$$\phi_Z(t) = p\phi_X(t) + (1-p)\phi_Y(t).$$

(c) **Inversion theorem:** If two random variables have the same characteristic function, then their distributions are the same. We prove this result below.

(d) The above inversion theorem remains valid for multivariate characteristic functions, defined by $\phi_{\mathbf{X}}(\mathbf{t}) = \mathbf{E}[e^{i\mathbf{t}^T\mathbf{X}}]$.

(e) For the univariate case, if X is a continuous random variable with PDF f_X, there is an explicit inversion formula, namely

$$f_X(x) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} e^{-itx} \phi_X(t)\, dt,$$

for every x at which f_X is differentiable. (Note the similarity with inversion formulas for Fourier transforms.)

(f) The dominated convergence theorem can be applied to complex random variables (simply apply the DCT separately to the complex and imaginary parts). Thus, if $\lim_{n\to\infty} X_n = X$, a.s., then, for every $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \phi_{X_n}(t) = \lim_{n\to\infty} \mathbf{E}[e^{itX_n}] = \mathbf{E}\big[\lim_{n\to\infty} e^{itX_n}\big] = \mathbf{E}[e^{itX}] = \phi_X(t).$$

The DCT applies here, because the random variables $|e^{itX_n}|$ are bounded by 1.

(g) If $\mathbf{E}[|X|^k] < \infty$, then $\phi_X(t)$ is k-times continuously differentiable and also

$$\frac{d^k}{dt^k}\phi_X(t)\Big|_{t=0} = i^k \mathbf{E}[X^k].$$

(This is plausible, by moving the differentiation inside the expectation, but a formal justification is needed.)

(h) If $\mathbf{E}[e^{\epsilon|X|}] < \infty$ for some $\epsilon > 0$ (equivalently if MGF of X exists in a neigborhood of zero) then $\phi_X(t)$ is analytic function of t, which extends to all complex z inside a strip $\{z : -\epsilon < \operatorname{Im} z < \epsilon\}$.

Two useful characteristic functions:

(a) **Exponential:** If $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$, then

$$\phi_X(y) = \frac{\lambda}{\lambda - it}.$$

Note that this is the same as starting with $M_X(s) = \lambda/(\lambda - s)$ and replacing s by it; however, this is not a valid proof. One must either use tools from complex analysis (contour integration), or evaluate separately $\mathbf{E}[\cos(tX)]$, $\mathbf{E}[\sin(tX)]$, which can be done using integration by parts.

(b) **Normal (scalar):** If $X \stackrel{d}{=} N(\mu, \sigma^2)$, then

$$\phi_X(t) = e^{it\mu} e^{-t^2\sigma^2/2}.$$

4.1 Inversion theorem

Theorem 3 (Inversion theorem)**.** *Let X and Y have the same characteristic functions. Then $\mathbb{P}_X = \mathbb{P}_Y$.*

Proof. Let $a > 1$ and consider the following "trapezoidal function"

$$f_a(x) = \begin{cases} 0, & |x| \geq a \\ \frac{1}{a-1}(x+a), & -a < x < -1 \\ 1, & -1 \leq x \leq 1 \\ -\frac{1}{a-1}(x-a), & 1 < x < a \end{cases}$$

Note that

$$\lim_{a \to 1+} f_a(x) = 1_{[-1,1]}(x) \tag{1}$$

Furthermore, there is an identity

$$f_a(x) = \frac{4}{(a-1)\sqrt{2\pi}} \int_{\mathbb{R}} e^{-itx} \frac{1}{t^2} \left[\frac{1}{a} \sin^2 \frac{ta}{2} - \sin^2 \frac{t}{2} \right] \, dt \tag{2}$$

To show this you may either compute the integral directly or use Fourier inversion and the observation that $f_a = \frac{1}{a-1}(g * g - h * h)$, where $g = 1_{[-a/2,a/2]}$, $h = 1_{[-1,1]}$ and $*$ is convolution.

Note that the integral in (2) is absolutely convergent since the absolute value of the integrand

$$\frac{1}{t^2} \left| \frac{1}{a} \sin^2 \frac{ta}{2} - \sin^2 \frac{t}{2} \right|$$

is continuous at 0 and integrable at $+\infty$. Thus, by Fubini we have

$$\mathbb{E}[f_a(X)] = \frac{4}{(a-1)\sqrt{2\pi}} \int_{\mathbb{R}} \phi_X(-t) \frac{1}{t^2} \left[\frac{1}{a} \sin^2 \frac{ta}{2} - \sin^2 \frac{t}{2} \right] \, dt$$

Since $\phi_X = \phi_Y$ we have

$$\mathbb{E}[f_a(X)] = \mathbb{E}[f_a(Y)]$$

for every $a > 1$. Taking limit as $a \searrow 1$ and applying the BCT to (1) we get

$$\mathbb{P}_X([-1,1]) = \mathbb{P}_Y([-1,1])$$

A similar argument (with shifted and scaled f_a) shows that $\mathbb{P}_X$ and $\mathbb{P}_Y$ coincide on every closed interval. Since the collection of closed intervals is a generating p-system, we have $\mathbb{P}_X = \mathbb{P}_Y$. $\square$

4.2 Vector-valued random variables

A very useful extension is to define characteristic function for vector-valued random variable $\mathbf{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$. In this case characteristic function is defined on $\mathbb{R}^d$ as follows:

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{X}}\right], \qquad \mathbf{t} = (t_1, \ldots, t_d)^T \in \mathbb{R}^d$$

where $\mathbf{t}^T\mathbf{X} = \sum_{j=1}^d t_j X_j$ denotes a standard scalar product on $\mathbb{R}^d$.

Most of the properties and results above (including inversion theorem) carry over to the vector case. This leads to numerous useful implications, of which we discuss two:

1. Checking independence: If $\mathbf{X} = (X_1, \ldots, X_d)^T$, then X_j are independent if and only if

   $$\phi_{\mathbf{X}}(\mathbf{t}) = \prod_{j=1}^{d} \phi_{X_j}(t_j) \tag{3}$$

   This easily follows from the inversion theorem, since right-hand side represents the characteristic function of distribution $\prod_{j=1}^d \mathbb{P}_{X_j}$.

2. Fourth definition of multivariate normal. It is not hard to show that for (degenerate or non-degenerate) multivariate normal $\mathbf{X}$ we have

   $$\phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mu^T\mathbf{t} - \frac{1}{2}\mathbf{t}^T V \mathbf{t}} \tag{4}$$

   where $\mu = \mathbf{E}[\mathbf{X}]$ and $V = \mathrm{Cov}(\mathbf{X}, \mathbf{X})$. Since ϕ uniquely determines the distribution, property (4) is frequently taken as the *definition* of a multivariate normal. Most properties then follow immediately. For example, "uncorrelated implies independent" is just a consequence of (3).