MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J Fall 2018
Lecture 14

MULTIVARIATE NORMAL DISTRIBUTIONS

Contents

In an earlier lecture, we worked through the bivariate normal distribution and its properties, relying mostly on algebraic manipulation and integration of normal PDFs. Here, we revisit the subject in more generality (n dimensions), while using more elegant tools. First, some background.

1 BACKGROUND ON POSITIVE DEFINITE MATRICES.

Definition 1. *Let A be a square ($n \times n$)* **symmetric** *matrix.*

(a) We say that A is **positive definite**, *and write $A > 0$, if $x^T Ax > 0$, for every nonzero $x \in \mathbb{R}^n$.*

(b) We say that A is **nonnegative definite**, *and write $A \geq 0$, if $x^T Ax \geq 0$, for every $x \in \mathbb{R}^n$.*

It is known (e.g., see any basic linear algebra text) that:

(a) A symmetric matrix has n real eigenvalues.

(b) A positive definite matrix has n real and positive eigenvalues.

(c) A nonnegative definite matrix has n real and nonnegative eigenvalues.

(d) To each eigenvalue of a symmetric matrix, we can associate a real eigenvector. Eigenvectors associated with distinct eigenvalues are orthogonal; eigenvectors associated with repeated eigenvalues can always be taken to be orthogonal. Without loss of generality, all these eigenvectors can be normalized so that they have unit length, resulting in an orthonormal basis.

(e) The above essentially states that a symmetric definite matrix becomes diagonal after a suitable orthogonal change of basis.

A concise summary of the above discussion is the following **spectral decomposition** formula: Every symmetric matrix A can be expressed in the form

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{z}_i \mathbf{z}_i^T,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is an associated collection of orthonormal eigenvectors. (Note here that $\mathbf{z}_i \mathbf{z}_i^T$ is a $n \times n$ matrix, of rank 1.)

For nonnegative definite matrices, we have $\lambda_i \geq 0$, which allows us to take square roots and define

$$B = \sum_{i=1}^{n} \sqrt{\lambda_i} \mathbf{z}_i \mathbf{z}_i^T.$$

We then observe that:

(a) The matrix B is symmetric.

(b) We have $B^2 = A$ (this is an easy calculation). Thus B is a **symmetric square root** of A.

(c) The matrix B has eigenvalues $\sqrt{\lambda_i}$. Therefore, it is positive (respectively, nonnegative) definite if and only if A is positive (respectively, nonnegative) definite.

Finally, if A is positive definite, then each λ_i is positive, and we can define the matrix

$$C = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{z}_i \mathbf{z}_i^T.$$

An easy calculation shows that $CA = AC = I$, so that $C = A^{-1}$.

2 DEFINITION OF THE MULTIVARIATE NORMAL DISTRIBUTION

Our interest in positive definite matrices stems for the following. When A is positive definite, the quadratic form $q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ goes to infinity as $\|\mathbf{x}\| \to \infty$, so that $e^{-q(\mathbf{x})}$ decays to zero, as $\|\mathbf{x}\| \to \infty$, and therefore can be used to define a multivariate PDF.

There are multiple ways of defining multivariate normal distributions. We will present three, and will eventually show that they are consistent with each other.

The first generalizes our definition of the bivariate normal. It is the most explicit and transparent; on the downside it can lead to unpleasant algebraic manipulations. Recall that $|V|$ stands for the absolute value of the determinant of a square matrix V.

Definition 2. *A random vector* $\mathbf{X}$ *has a* **nondegenerate (multivariate) normal** *distribution if it has a joint PDF of the form*

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

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

The second definition is constructive, which makes it operationally useful.

Definition 3. *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.*

The last definition is possibly the hardest to penetrate, but in the eyes of some, it is the most elegant.

Definition 4. *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.*

A brief remark on the use of the word "nondegenerate" in Definition 2. Under Definition 2, $f_X(\mathbf{x}) > 0$ for all $\mathbf{x}$. On the other hand, consider the following example. Let $X_1 \sim N(0,1)$ and let $X_2 = 0$. The random vector $\mathbf{X} = (X_1, X_2)$ is normal according to Definitions 3 or 4, but cannot be described by a joint PDF (all of the probability is concentrated on the horizontal axis, a set of zero area). This is an example of a degenerate normal distribution: the distribution is concentrated on a proper subspace of $\mathbb{R}^n$. The most extreme example is a one-dimensional random variable, which is identically equal to zero. This qualifies as normal under Definitions 3 and 4. One may question the wisdom of calling the number "zero" a "normal random variable;" the reason for doing so is that it allows us to state results such as "a linear function of a normal random variable is normal", etc., without having to worry about exceptions and special conditions that will prevent degeneracy.

3 MEANS AND COVARIANCES OF VECTOR RANDOM VARIABLES

Let us first introduce a bit more notation. If $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector, we define

$$\mathbf{E}[\mathbf{X}] = (\mathbf{E}[X_1], \ldots, \mathbf{E}[X_n]),$$

which we treat as a column vector. Similarly, If $\mathbf{A}$ is a random matrix (a matrix with each entry being a random variable A_{ij}), we use the notation $\mathbf{E}[\mathbf{A}]$, to denote the matrix whose entries are $\mathbf{E}[A_{ij}]$.

Given two random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$, we can consider all the possible covariances

$$\mathrm{Cov}(X_i, Y_j) = \mathbf{E}[(X_i - \mathbf{E}[X_i])(Y_i - \mathbf{E}[Y_j])],$$

and we can arrange them in a $n \times m$ **covariance matrix**

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$$

whose (i,j)th entry is $\mathrm{Cov}(X_i, Y_j)$. It is easily checked that

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{E}\Big[(X - \mathbf{E}[X])(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])^T\Big].$$

Notice also that $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$ is a $n \times n$ symmetric matrix.

Exercise 1. Prove that $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$ is nonnegative definite.

4 KEY PROPERTIES OF THE MULTIVARIATE NORMAL

The theorem below includes almost everything useful there is to know about multivariate normals. We will prove and state the theorem, while working mostly with Definition 3. The proof of equivalence of the three definitions will be completed in the next lecture, together with some additional observations.

Theorem 1. *Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ is multivariate normal, in the sense of Definition 3, and let μ_i be the ith component of μ.*

(a) For every i, X_i is normal, with mean μ_i.

(b) We have $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = DD^T$.

(c) If C is a $m \times n$ matrix and $\mathbf{d}$ is a vector in $\mathbb{R}^m$, then $Y = C\mathbf{X} + d$ is multivariate normal in the sense of Definition 3, with mean $C\mu + \mathbf{d}$ and covariance matrix CDD^TC^T.

(d) If $|D| \neq 0$, then $\mathbf{X}$ is a nondegenerate multivariate normal in the sense of Definition 2, with $V = DD^T = \mathrm{Cov}(\mathbf{X}, \mathbf{X})$.

(e) The joint CDF F_X of $\mathbf{X}$ is completely determined by the mean and covariance of $\mathbf{X}$.

(f) The components of $\mathbf{X}$ are uncorrelated (i.e., the covariance matrix is diagonal) if and only if they are independent.

(g) If

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix} \right),$$

and $V_{YY} > 0$, then:

(i) $\mathbf{E}[\mathbf{X} \mid \mathbf{Y}] = \mu_X + V_{XY}V_{YY}^{-1}(\mathbf{Y} - \mu_Y)$.

(ii) Let $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{E}[\mathbf{X} \mid \mathbf{Y}]$. Then, $\tilde{\mathbf{X}}$ is independent of $\mathbf{Y}$, and independent of $\mathbf{E}[\mathbf{X} \mid \mathbf{Y}]$.

(iii) $\mathrm{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}} \mid \mathbf{Y}) = \mathrm{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) = V_{XX} - V_{XY}V_{YY}^{-1}V_{YX}$.

Proof:

(a) Under definition 3, X_i is a linear function of independent normal random variables, hence normal. Since $\mathbf{E}[\mathbf{W}] = 0$, we have $\mathbf{E}[X_i] = \mu_i$.

(b) For simplicity, let us just consider the zero mean case. We have

$$\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathbf{E}[\mathbf{X}\mathbf{X}^T] = \mathbf{E}[D\mathbf{W}\mathbf{W}^TD^T] = D\mathbf{E}[\mathbf{W}\mathbf{W}^T]D^T = DD^T,$$

where the last equality follows because the components of $\mathbf{W}$ are independent (hence the covariance matrix is diagonal), with unit variance (hence the diagonal entries are all equal to 1).

(c) We have $\mathbf{Y} = C\mathbf{X} + d = C(D\mathbf{W} + \mu) + d$, which is itself a linear function of independent standard normal random variables. Thus, $\mathbf{Y}$ is multivariate normal. The formula for $\mathbf{E}[\mathbf{Y}]$ is immediate. The formula for the covariance matrix follows from part (b), with D being replaced by (CD).

(d) This is an exercise in derived distributions. Let us again just consider the case of $\mu = 0$. We already know (Lecture 10) that when $X = DW$, with D invertible, then

$$f_X(x) = \frac{f_W(D^{-1}x)}{|\det D|}.$$

In our case, since the W_i are i.i.d. N(0,1), we have

$$f_W(\mathbf{w}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\Big\{ -\frac{1}{2}\mathbf{w}^T\mathbf{w}\Big\},$$

leading to

$$f_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |DD^T|}} \exp\Big\{ -\frac{1}{2}\mathbf{x}^T(D^{-1})^T D^{-1}\mathbf{x}\Big\}.$$

This is of the form given in Definition 2, with $V = DD^T$. In conjunction with part (b), we also have $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = V$. The argument for the nonzero mean case is essentially the same.

(e) Using part (d), the joint PDF of $\mathbf{X}$ is completely determined by the matrix V, which happens to be equal to $\mathrm{Cov}(\mathbf{X}, \mathbf{X})$, together with the vector μ.

The degenerate case is a little harder, because of the absence of a convenient closed form formula. One could think of a limiting argument that involves injecting a tiny bit of noise in all directions, to make the distribution nondegenerate, and then taking the limit. This type of argument can be made to work, but will involve tedious technicalities. Instead, we will take a shortcut, based on the inversion property of transforms. This argument is simpler, but relies on the heavy machinery behind the proof of the inversion property.

Let us find the multivariate transform $M_{\mathbf{X}}(\mathbf{s}) = \mathbf{E}[e^{\mathbf{s}^T\mathbf{x}}]$. We note that $\mathbf{s}^T\mathbf{X}$ is normal with mean $\mathbf{s}^T\mu$. Letting $\tilde{\mathbf{X}} = \mathbf{X} - \mu$, the variance of $\mathbf{s}^T\mathbf{X}$

is

$$\mathrm{var}(\mathbf{s}^T\mathbf{X}) = \mathbf{E}[\mathbf{s}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{s}] = \mathbf{s}^T\mathbf{E}[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T]\mathbf{s} = \mathbf{s}^T\mathrm{Cov}(\mathbf{X},\mathbf{X})\mathbf{s} = \mathbf{s}^T V\mathbf{s}.$$

Using the formula for the transform of a single normal random variable ($\mathbf{s}^T\mathbf{X}$ in this case), we have

$$M_{\mathbf{X}}(\mathbf{s}) = \mathbf{E}[e^{\mathbf{s}^T\mathbf{x}}] = M_{\mathbf{s}^T\mathbf{X}}(1) = e^{\mathbf{s}^T\mu}e^{\mathbf{s}^T V\mathbf{s}/2}.$$

Thus, μ and V completely determine the transform of $\mathbf{X}$. By the inversion property of transforms, μ and V completely determine the distribution (e.g., the CDF) of $\mathbf{X}$.

(f) If the components of $\mathbf{X}$ are independent they are of course uncorrelated. For the converse, suppose that the components of $\mathbf{X}$ are uncorrelated, i.e., the matrix V is a diagonal. Consider another random vector $\mathbf{Y}$ that has the same mean and as $\mathbf{X}$, whose components are independent normal, and such that the variance of Y_i is the same as the variance of X_i. Then, $\mathbf{X}$ and $\mathbf{Y}$ have the same mean and covariance. By part (e), $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution. Since the components of $\mathbf{Y}$ are independent, it follows that the components of $\mathbf{X}$ are also independent.

For the special case when V is invertible, we could alternatively use part (d) which provides an explicit formula for the joint PDF of $\mathbf{X}$. When V is diagonal we see that the joint PDF is the product of its marginal PDF. Namely $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}_1}(\mathbf{x}_1)\cdots f_{\mathbf{X}_n}(\mathbf{x}_n)$ and thus the components of $\mathbf{X}$ are independent.

(g) Once more, to simplify notation, let us just deal with the zero-mean case. Let us define

$$\hat{\mathbf{X}} = V_{XY}V_{YY}^{-1}\mathbf{Y}.$$

We then have

$$\mathbf{E}[\hat{\mathbf{X}}\mathbf{Y}^T] = V_{XY}V_{YY}^{-1}\mathbf{E}[\mathbf{Y}\mathbf{Y}^T] = V_{XY} = \mathbf{E}[\mathbf{X}\mathbf{Y}^T].$$

This proves that $\mathbf{X} - \hat{\mathbf{X}}$ is uncorrelated with $\mathbf{Y}$. Note that $(\mathbf{X} - \hat{\mathbf{X}}, \mathbf{Y})$ is a linear function of $(\mathbf{X}, \mathbf{Y})$, so, by part (c), it is also multivariate normal. Using an argument similar to the one in the proof of part (f), we conclude that $\mathbf{X} - \hat{\mathbf{X}}$ is independent of $\mathbf{Y}$, and therefore independent from any function of $\mathbf{Y}$. Recall now the abstract definition of conditional expectations. The relation $\mathbf{E}[(\mathbf{X} - \hat{\mathbf{X}})g(\mathbf{Y})] = 0$, for every function g, implies that $\hat{\mathbf{X}} = \mathbf{E}[\mathbf{X} \mid \mathbf{Y}]$, which proves part (i).

For part (ii), note that we already proved that $\tilde{\mathbf{X}} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{E}[\mathbf{X} \mid \mathbf{Y}]$ is independent of $\mathbf{Y}$. Since $\mathbf{E}[\mathbf{X} \mid \mathbf{Y}]$ is a function of $\mathbf{Y}$, it follows that $\tilde{\mathbf{X}}$ is independent of $\mathbf{E}[\mathbf{X} \mid \mathbf{Y}]$.

For part (iii), note that $\tilde{\mathbf{X}}$ is independent of $\mathbf{Y}$, which implies that $\mathrm{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}} \mid \mathbf{Y}) = \mathrm{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}})$. Finally,

$$\begin{aligned}\mathrm{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) &= \mathbf{E}[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T] = \mathbf{E}[(\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})^T] = \mathbf{E}[(\mathbf{X} - \hat{\mathbf{X}})\mathbf{X}^T] \\ &= V_{XX} - \mathbf{E}[V_{XY}V_{YY}^{-1}\mathbf{Y}\mathbf{X}^T] = V_{XX} - V_{XY}V_{YY}^{-1}\mathbf{E}[\mathbf{Y}\mathbf{X}^T] \\ &= V_{XX} - V_{XY}V_{YY}^{-1}V_{YX}.\end{aligned}$$

□

Note that in the case of the bivariate normal, we have $\mathrm{Cov}(X, Y) = \rho\sigma_X\sigma_Y$, $V_{XX} = \sigma_X^2$, $V_{YY} = \sigma_Y^2$. Then, part (g) of the preceding theorem, for the zero-mean case, reduces to

$$\mathbf{E}[X \mid Y] = \rho\frac{\sigma_X}{\sigma_Y}Y, \qquad \mathrm{var}(\tilde{X}) = \sigma_X^2(1 - \rho^2),$$

which agrees with the formula we derived through elementary means in Lecture 9, for the special case of unit variances.

6.436J / 15.085J Fundamentals of Probability

Fall 2018