Multivariate Distributions and Associated Differential Entropy of Jointly Normal Random Variables

H.L. Rappaport

October 6, 2014

In this note, the multivariate normal distribution of real random variables is reviewed in some detail. The treatment is that of Hogg and Craig [1]. The differential entropy of these distributions will also be found.

Consider a real symmetric $n \times n$ matrix $\mathbf{A}$ which is positive definite. A real symmetric matrix is positive definite if the quadratic form [2]

$$Q = \mathbf{x}^T \mathbf{A} \mathbf{x}, \tag{1}$$

satisfies

$$Q \geq 0, \tag{2}$$

for all real column vectors $\mathbf{x}$ *and* $Q = 0$ *only* for $\mathbf{x} = \mathbf{0}$, e.g., $\mathbf{x}$ is zero for all elements. The computations make use of lower case boldface letters for column vectors, e.g., $\mathbf{x}$, $\boldsymbol{\mu}$, $\mathbf{y}$, $\mathbf{t}$, $\mathbf{w}$ where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}, \tag{3}$$

etc., and upper case boldface letters for matrices.

A joint p.d.f. of n real continuous type random variables $X_1, X_2, \ldots, X_n$ is considered

$$f(\mathbf{x}) = C \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{x} - \boldsymbol{\mu})\right], \tag{4}$$

where $\boldsymbol{\mu}$ is a constant vector and the constant C is to be determined. The moment-generating function for $f(\mathbf{x})$ is given by

$$M(\mathbf{t}) = C \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\mathbf{t}^T \mathbf{x} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{x} - \mathbf{u})\right] dx_1 \cdots dx_n. \tag{5}$$

or

$$M(\mathbf{t}) = C \int \exp\left[\mathbf{t}^T \mathbf{x} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{x} - \mathbf{u})\right] d^n x, \tag{6}$$

to simplify notation.

Changing variables under the integrals in Eq. (5) to $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$ produces

$$M(\mathbf{t}) = C \exp\left(\mathbf{t}^T \boldsymbol{\mu}\right) \int \exp\left(\mathbf{t}^T \mathbf{y} - \frac{1}{2}\mathbf{y}^T \mathbf{A}\mathbf{y}\right) d^n y. \tag{7}$$

Now every real symmetric matrix is orthogonally similar to a diagonal matrix (Wylie [2], p. 553). In other words, there exists a matrix $\mathbf{L}$ that diagonalizes $\mathbf{A}$ via a similarity transformation, e.g.,

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{L} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}, \tag{8}$$

where the $\{\lambda_i\}$ are the eigenvalues of $\mathbf{A}$. Since the matrix is orthogonal $\mathbf{L}^T = \mathbf{L}^{-1}$. It will be convenient to introduce the notation $\text{diag}\left[\lambda_1, \lambda_2, \ldots, \lambda_n\right]$ to denote a diagonal matrix such as appearing in Eq. (8).

The next step is a further change of variables to $\mathbf{z}$ via $\mathbf{y} = \mathbf{L}\mathbf{z}$. The determinant of a real orthogonal matrix is ± 1. To see this note that

$$\left|\mathbf{L}^T \mathbf{L}\right| = 1, \tag{9}$$

where $|\cdot|$ denotes determinant, since $\mathbf{L}^T = \mathbf{L}^{-1}$. The determinant of a product of two matrices equals the product of the determinants of each matrix. Also, the determinant of the transpose of a matrix is equal to the determinant of the matrix [3]. So Eq. (9) can be written

$$|\mathbf{L}|^2 = 1, \tag{10}$$

and the determinant is ± 1 as stated. Therefore, the absolute value of the Jacobian of the transformation $\mathbf{y} = \mathbf{L}\mathbf{z}$ is $+1$ and $d^n z = d^n y$ may be used in the integral of Eq. (7) which now becomes

$$M(\mathbf{t}) = C \exp\left(\mathbf{t}^T \boldsymbol{\mu}\right) \int \exp\left(\mathbf{t}^T \mathbf{L}\mathbf{z} - \frac{1}{2}\mathbf{z}^T \mathbf{L}^T \mathbf{A}\mathbf{L}\mathbf{z}\right) d^n z. \tag{11}$$

Now $\mathbf{w} = \mathbf{L}^T \mathbf{t}$ or equivalently $\mathbf{t} = \mathbf{L}\mathbf{w}$ is introduced under the integral. Also,

$$\mathbf{z}^T \mathbf{L}^T \mathbf{A}\mathbf{L}\mathbf{z} = \sum_{i=1}^{n} \lambda_i z_i^2, \tag{12}$$

from Eq. (8). In this way, Eq. (7) becomes

$$M(\mathbf{t}) = C\exp\left(\mathbf{t}^T\boldsymbol{\mu}\right)\int \exp\left[\sum_{i=1}^{n}\left(w_i z_i - \frac{1}{2}\lambda_i z_i^2\right)\right] d^n z, \tag{13}$$

$$= C\exp\left(\mathbf{t}^T\boldsymbol{\mu}\right)\prod_{i=1}^{n}\int_{-\infty}^{\infty}\exp\left(w_i z_i - \frac{1}{2}\lambda_i z_i^2\right) dz_i. \tag{14}$$

Now

$$w_i z_i - \frac{1}{2}\lambda_i z_i^2 = \frac{w_i^2}{2\lambda_i} - \frac{\lambda_i}{2}\left(z_i - \frac{w_i}{\lambda_i}\right)^2, \tag{15}$$

and the integral appearing in Eq. (14) is easily evaluated leading to

$$M(\mathbf{t}) = C\exp\left(\mathbf{t}^T\boldsymbol{\mu}\right)\prod_{i=1}^{n}\exp\left(\frac{w_i^2}{2\lambda_i}\right)\sqrt{\frac{2\pi}{\lambda_i}}, \tag{16}$$

$$= C\exp\left(\mathbf{t}^T\boldsymbol{\mu}\right)\sqrt{\frac{(2\pi)^n}{\lambda_1\lambda_2\ldots\lambda_n}}\exp\left(\sum_{i=1}^{n}\frac{w_i^2}{2\lambda_i}\right). \tag{17}$$

In obtaining these results $\lambda_i > 0$ has been used in order that the integrals appearing in Eq. (14) converge. This is justified because a real symmetric positive definite matrix has positive eigenvalues [4].

Next an intermediate result is derived. In light of Eq. (8)

$$\left(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}\right)^{-1} = \mathrm{diag}\left[\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots \frac{1}{\lambda_n}\right], \tag{18}$$

but

$$\left(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}\right)^{-1} = \mathbf{L}^{-1}\mathbf{A}^{-1}\mathbf{L} = \mathbf{L}^T\mathbf{A}^{-1}\mathbf{L}, \tag{19}$$

so

$$\mathbf{L}^T\mathbf{A}^{-1}\mathbf{L} = \mathrm{diag}\left[\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots \frac{1}{\lambda_n}\right], \tag{20}$$

the required result.

Using Eq. (20)

$$\sum_{i=1}^{n}\frac{w_i^2}{\lambda_i} = \mathbf{w}^T\left(\mathbf{L}^T\mathbf{A}^{-1}\mathbf{L}\right)\mathbf{w} = \left(\mathbf{L}\mathbf{w}\right)^T\mathbf{A}^{-1}\left(\mathbf{L}\mathbf{w}\right) = \mathbf{t}^T\mathbf{A}^{-1}\mathbf{t}. \tag{21}$$

Also, the determinant of $\mathbf{A}^{-1}$ is

$$\left|\mathbf{A}^{-1}\right| = \left|\mathbf{L}^T\mathbf{A}^{-1}\mathbf{L}\right| = \frac{1}{\lambda_1\lambda_2\cdots\lambda_n}. \tag{22}$$

Thus Eq. (17) becomes

$$M(\mathbf{t}) = C\sqrt{(2\pi)^n\left|\mathbf{A}^{-1}\right|}\exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\mathbf{A}^{-1}\mathbf{t}\right). \tag{23}$$

Having obtained the result of Eq. (23) the quantity C can now be evaluated. The normalization of the density function requires

$$M(\mathbf{t}=\mathbf{0}) = \int f(\mathbf{x})\; d^n x = 1, \tag{24}$$

or equivalently

$$M(\mathbf{t}=\mathbf{0}) = C\sqrt{(2\pi)^n\left|\mathbf{A}^{-1}\right|} = 1. \tag{25}$$

So

$$C = \frac{1}{\sqrt{(2\pi)^n\left|\mathbf{A}^{-1}\right|}}, \tag{26}$$

and Eq. (4) becomes

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n\left|\mathbf{A}^{-1}\right|}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{27}$$

and Eq. (23) becomes

$$M(\mathbf{t}) = \exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\mathbf{A}^{-1}\mathbf{t}\right). \tag{28}$$

Let the elements of $\mathbf{A}^{-1}$ be denoted by σ_{ij}; $\;i,j \in \{1,2,\ldots,n\}$. The inverse of a symmetric matrix, if it exists, is a symmetric matrix, e.g.,

$$\left(\mathbf{A}^{-1}\right)^T = \left(\mathbf{A}^T\right)^{-1} = \mathbf{A}^{-1}, \tag{29}$$

so the elements of $\mathbf{A}^{-1}$ satisfy $\sigma_{ij} = \sigma_{ji}$.

From Eq. (28)

$$M(\mathbf{t}) = \int \exp\left(\sum_{i=1}^n t_i x_i\right) f(\mathbf{x})\; d^n x = \exp\left(\sum_{i=1}^n t_i\mu_i + \frac{1}{2}\sum_{i,j=1}^n t_i t_j\sigma_{ij}\right). \tag{30}$$

If $t_i = 0,\ i \neq k$ then

$$M = \exp\left(t_k\mu_k + \frac{1}{2}t_k^2\sigma_{kk}\right), \tag{31}$$

which is the moment-generating function for a normal distribution with mean μ_k and variance σ_{kk} [5]. Furthermore,

If $t_i = 0,\ i \neq k, l$ then

$$M = \exp\left[t_k\mu_k + t_l\mu_l + \frac{1}{2}\left(t_k^2\sigma_{kk} + 2t_kt_l\sigma_{kl} + t_l^2\sigma_{ll}\right)\right], \tag{32}$$

which is the moment-generating function of a bivariate normal distribution with mean μ_k and μ_l and variances σ_{kk} and σ_{ll} and covariance σ_{kl} [5]. The importance of these results is emphasized with the following discussion of characteristic functions.

Characteristic Functions

The characteristic function of a random variable X is defined (Papoulis [6], p. 153)

$$\Phi(t) = \int_{-\infty}^{\infty} \exp\left(\hat{i}tx\right) f_X(x)\ dx. \tag{33}$$

In other words, the characteristic function is the Fourier transform of the distribution function. Note that $\hat{i}$ is equal to $\sqrt{-1}$.

Since

$$M(t) = \int_{-\infty}^{\infty} \exp(tx)\, f_X(x)\ dx, \tag{34}$$

it is seen that

$$\Phi(t) = M\left(\hat{i}t\right), \tag{35}$$

for values of t where both functions exist. Thus, from Eq. (31), the normal p.d.f. has characteristic function

$$\Phi(t) = \exp\left(\hat{i}\mu_x t - \frac{1}{2}\sigma_x^2 t^2\right), \tag{36}$$

in agreement with Papoulis [6] p. 159.

The joint characteristic function of two random variables (Papoulis, p. 213) X and Y is defined

$$\Phi(t_x, t_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[\hat{i}\left(t_x x + t_y y\right)\right] f_{XY}(x, y)\ dx\, dy. \tag{37}$$

In other words, the joint characteristic function of two random variables is the double Fourier transform of the joint distribution function. Since

$$M\left(t_x, t_y\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(t_x x + t_y y\right) f_{XY}\left(x, y\right)\, dx\, dy, \tag{38}$$

it is seen that

$$\Phi\left(t_x, t_y\right) = M\left(\hat{i}t_x, \hat{i}t_y\right), \tag{39}$$

for values of t_x and t_y where both functions exist. Thus from prior work [5] the p.d.f. for two jointly normal random variables has the joint characteristic function

$$\Phi\left(t_x, t_y\right) = \exp\left[\hat{i}\left(\mu_x t_x + \mu_y t_y\right) - \frac{1}{2}\left(t_x^2\sigma_x^2 + 2r t_x t_y \sigma_x \sigma_y + t_y^2\sigma_y^2\right)\right], \tag{40}$$

in agreement with Papoulis p. 226.

Finally, the joint characteristic function of n random variables is given by (Papoulis, p. 244)

$$\Phi\left(\mathbf{t}\right) = \int \exp\left(\hat{i}\mathbf{t}^T\mathbf{x}\right) f\left(\mathbf{x}\right)\, d^n\mathbf{x}, \tag{41}$$

while the moment-generating function for n random variables satisfies

$$M\left(\mathbf{t}\right) = \int \exp\left(\mathbf{t}^T\mathbf{x}\right) f\left(\mathbf{x}\right)\, d^n\mathbf{x}, \tag{42}$$

so

$$\Phi\left(\mathbf{t}\right) = M\left(\hat{i}\mathbf{t}\right). \tag{43}$$

Thus the joint characteristic function of n random variables follows from Eq. (28), e.g.,

$$\Phi\left(\mathbf{t}\right) = \exp\left(\hat{i}\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\mathbf{A}^{-1}\mathbf{t}\right), \tag{44}$$

in agreement with Papoulis p. 255.

Thus the conclusion that setting $t_i = 0$, $i \neq k$ or setting $t_i = 0$, $i \neq k, l$ reduces the moment-generating function of the p.d.f. given in Eq. (4) to the moment-generating function of the normal distribution and the bivariate jointly normal distribution respectively carries over to the characteristic functions. That is, setting $t_i = 0$, $i \neq k$ or setting $t_i = 0$, $i \neq k, l$ reduces the characteristic function of the p.d.f. of Eq. (4) to the characteristic function of the normal distribution and the bivariate jointly normal distributions respectively.

From the definition of the characteristic function, Eq. (41), or the definition of the moment-generating function, Eq. (42), it is clear that setting several of the t_i to zero in either function is equivalent to evaluating the characteristic function or moment-generating function of the corresponding marginal density or joint density of the remaining X_i, e.g., the X_i conjugate to the retained t_i.

The relationships just discussed show that the elements of $\boldsymbol{\mu}$ are the means of the X_i, the elements of $\mathbf{A}^{-1}$ on the principal diagonal are the variances $\sigma_{ii} = \sigma_i^2$ of the X_i, and the off-diagonal elements of $\mathbf{A}^{-1}$ or σ_{ij}, $i \neq j$, are the covariances of X_i and X_j.

For these reasons, the matrix $\mathbf{A}^{-1}$ is called the covariance matrix of the multivariate normal distribution. Since the inverse matrix is also normal and symmetric and has positive eigenvalues it is also positive definite [4]. Denoting $\mathbf{A}^{-1}$ by $\mathbf{V}$ we have from Eqs. (27) and (28)

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{45}$$

and

$$M(\mathbf{t}) = \exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\mathbf{V}\mathbf{t}\right), \tag{46}$$

as given by Hogg and Craig.

From a further perspective

$$E[(X_k - \mu_k)(X_l - \mu_l)] = E(X_k X_l) - \mu_k\mu_l, \tag{47}$$

and using Eq. (30)

$$E[(X_k - \mu_k)(X_l - \mu_l)] = \left.\frac{\partial^2 M}{\partial t_k \partial t_l}\right|_{\mathbf{t}=\mathbf{0}} - \mu_k\mu_l. \tag{48}$$

Also from Eq. (30)

$$\left.\frac{\partial^2 M}{\partial t_k \partial t_l}\right|_{\mathbf{t}=\mathbf{0}} = \left.\frac{\partial}{\partial t_l}\right|_{\mathbf{t}=\mathbf{0}} \left[\left(\mu_k + \sum_{j=1}^{n} t_j\sigma_{jk}\right) M(\mathbf{t})\right], \tag{49}$$

$$= \left[\sigma_{kl}M(\mathbf{t}) + \left(\mu_k + \sum_{j=1}^{n} t_j\sigma_{jk}\right)\left(\mu_l + \sum_{i=1}^{n} t_i\sigma_{il}\right) M(\mathbf{t})\right]_{\mathbf{t}=\mathbf{0}}, \tag{50}$$

$$= \sigma_{kl} + \mu_k\mu_l, \tag{51}$$

so

$$E\left[\left(X_k-\mu_k\right)\left(X_l-\mu_l\right)\right]=\sigma_{kl}, \tag{52}$$

or

$$E\left[\left(\mathbf{X}-\boldsymbol{\mu}\right)\left(\mathbf{X}^T-\boldsymbol{\mu}^T\right)\right]=E\left(\mathbf{X}\mathbf{X}^T\right)-\boldsymbol{\mu}\boldsymbol{\mu}^T=\mathbf{V}, \tag{53}$$

displaying the significance of the covariance matrix in vector form.

Differential Entropy of Multivariate Normal Distributions

In this section, the differential entropy of a multivariate distribution of jointly normal real random variables is found. The presentation follows that given by Cover and Thomas [7]. The joint differential entropy of n random variables satisfies

$$h\left(\mathbf{X}\right)=-\int f\left(\mathbf{x}\right)\log_2 f\left(\mathbf{x}\right)\,d^nx. \tag{54}$$

For the distribution given in Eq. (45)

$$h\left(\mathbf{X}\right)=-\frac{1}{\ln 2}\int f\left(\mathbf{x}\right)\left\{-\frac{1}{2}\ln\left[\left(2\pi\right)^n\left|\mathbf{V}\right|\right]-\frac{1}{2}\left[\left(\mathbf{x}-\boldsymbol{\mu}\right)^T\mathbf{V}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]\right\}d^nx, \tag{55}$$

$$=\frac{1}{2}\log_2\left[\left(2\pi\right)^n\left|\mathbf{V}\right|\right]+\frac{1}{2\ln 2}E\left[\left(\mathbf{X}-\boldsymbol{\mu}\right)^T\mathbf{V}^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\right)\right]. \tag{56}$$

$$=\frac{1}{2}\log_2\left[\left(2\pi\right)^n\left|\mathbf{V}\right|\right]+\frac{1}{2\ln 2}E\left[\sum_{i,j=1}^{n}\left(X_i-\mu_i\right)\left(\mathbf{V}^{-1}\right)_{ij}\left(X_j-\mu_j\right)\right], \tag{57}$$

$$=\frac{1}{2}\log_2\left[\left(2\pi\right)^n\left|\mathbf{V}\right|\right]+\frac{1}{2\ln 2}\sum_{i,j=1}^{n}E\left[\left(X_i-\mu_i\right)\left(\mathbf{V}^{-1}\right)_{ij}\left(X_j-\mu_j\right)\right], \tag{58}$$

or

$$h\left(\mathbf{X}\right)=\frac{1}{2}\log_2\left[\left(2\pi\right)^n\left|\mathbf{V}\right|\right]+\frac{1}{2\ln 2}\sum_{i,j=1}^{n}\left(\mathbf{V}^{-1}\right)_{ij}E\left[\left(X_i-\mu_i\right)\left(X_j-\mu_j\right)\right]. \tag{59}$$

From Eq. (52) this becomes

$$h\left(\mathbf{x}\right)=\frac{1}{2}\log_2\left[\left(2\pi\right)^n\left|\mathbf{V}\right|\right]+\frac{1}{2\ln 2}\sum_{i,j=1}^{n}\left(\mathbf{V}^{-1}\right)_{ij}\left(\mathbf{V}\right)_{ij}, \tag{60}$$

$$=\frac{1}{2}\log_2\left[\left(2\pi\right)^n\left|\mathbf{V}\right|\right]+\frac{1}{2\ln 2}\sum_{j=1}^{n}\delta_{jj}, \tag{61}$$

$$= \frac{1}{2} \log_2 \left[(2\pi)^n \left| \mathbf{V} \right| \right] + \frac{1}{2} \log_2 e^n, \tag{62}$$

$$= \frac{1}{2} \log_2 \left[(2\pi e)^n \left| \mathbf{V} \right| \right], \tag{63}$$

the desired result.

References

[1] R. V. Hogg and A.T. Craig, *Introduction to Mathematical Statistics*, Macmillan, N.Y. (1978), Chap. 12.

[2] C. R. Wylie, *Advanced Engineering Mathematics, Fourth Ed.*, McGraw-Hill, N.Y. (1960), Chap. 11.

[3] I. Reiner, *Introduction to Matrix Theory and Linear Algebra*, Holt, Rinehart and Winston, N.Y. (1971), p. 18.

[4] B. Kolman, *Elementary Linear Algebra*, Macmillan, N.Y. (1977), p. 264.

[5] H. L. Rappaport, *Normal and Bivariate Normal Distributions and Moment-Generating Functions*, 7G Communications, 7GCTN03, October 2014.

[6] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, N.Y. (1965),

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, N.Y. (1991), p. 230.