Multivariate normal distribution

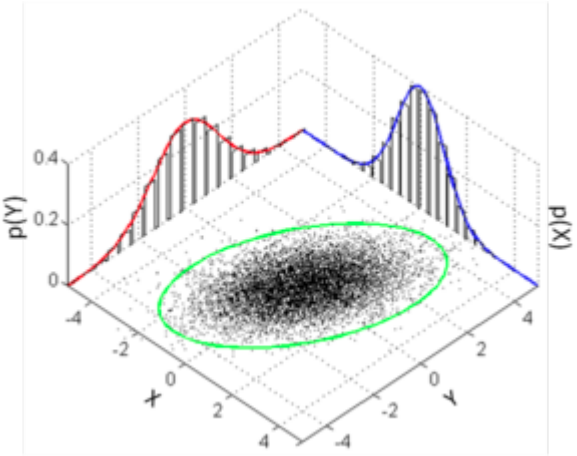
From Wikipedia, the free encyclopedia

In probability theory and statistics, the **multivariate normal distribution** or **multivariate Gaussian distribution**, is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. One possible definition is that a random vector is said to be *k*-variate normally distributed if every linear combination of its *k* components has a univariate normal distribution. Its importance derives mainly from the multivariate central limit theorem. The multivariate normal distribution is often used to describe, at least approximately, any set of (possibly) correlated real-valued random variables each of which clusters around a mean value.

Contents

- 1 Notation and parametrization
- 2 Definition
- 3 Properties
  - 3.1 Density function
    - 3.1.1 Non-degenerate case
    - 3.1.2 Degenerate case
  - 3.2 Higher moments
  - 3.3 Likelihood function
  - 3.4 Entropy
  - 3.5 Kullback–Leibler divergence
  - 3.6 Cumulative distribution function
  - 3.7 Prediction Interval
- 4 Joint normality
  - 4.1 Normally distributed and independent
  - 4.2 Two normally distributed random variables need not be jointly bivariate normal
  - 4.3 Correlations and independence
- 5 Conditional distributions

Probability density function



Many samples from a multivariate normal distribution, shown along with the 3-sigma ellipse, the two marginal distributions, and the two 1-d histograms.

Notation	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Parameters	$\boldsymbol{\mu} \in \mathbf{R}^k$ — location $\boldsymbol{\Sigma} \in \mathbf{R}^{k \times k}$ — covariance (nonnegative-definite matrix)
Support	$\boldsymbol{x} \in \boldsymbol{\mu}+\text{span}(\boldsymbol{\Sigma}) \subseteq \mathbf{R}^k$
PDF	$(2\pi)^{-\frac{k}{2}}\lvert\boldsymbol{\Sigma}\rvert^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$, exists only when $\boldsymbol{\Sigma}$ is positive-definite
CDF	(no analytic expression)
Mean	$\boldsymbol{\mu}$
Mode	$\boldsymbol{\mu}$
Variance	$\boldsymbol{\Sigma}$
Entropy	$\frac{k}{2}(1+\ln(2\pi)) + \frac{1}{2}\ln\lvert\boldsymbol{\Sigma}\rvert$
MGF	$\exp\left(\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right)$
CF	$\exp\left(i\boldsymbol{\mu}'\mathbf{t} - \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right)$

- 5.1 Bivariate case
- 5.2 Bivariate conditional expectation
  - 5.2.1 In the general case
  - 5.2.2 In the case of unit variances
- 6 Marginal distributions
- 7 Affine transformation
- 8 Geometric interpretation
- 9 Estimation of parameters
- 10 Bayesian inference
- 11 Multivariate normality tests
- 12 Drawing values from the distribution
- 13 See also
- 14 References
  - 14.1 Literature

Notation and parametrization

The multivariate normal distribution of a k-dimensional random vector $\mathbf{x} = [X_1, X_2, \ldots, X_k]$ can be written in the following notation:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

or to make it explicitly known that X is k-dimensional,

$$\mathbf{x} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

with k-dimensional mean vector

$$\boldsymbol{\mu} = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]$$

and $k \times k$ covariance matrix

$$\boldsymbol{\Sigma} = [\mathrm{Cov}[X_i, X_j]], i = 1, 2, \ldots, k; j = 1, 2, \ldots, k$$

Definition

A random vector $\mathbf{x} = (X_1, \ldots, X_k)'$ is said to have the multivariate normal distribution if it satisfies the following equivalent conditions.[1]

- Every linear combination of its components $Y = a_1X_1 + \ldots + a_kX_k$ is normally distributed. That is, for any constant vector $\mathbf{a} \in \mathbf{R}^k$, the random variable $Y = \mathbf{a}'\mathbf{x}$ has a univariate normal distribution, where a univariate normal distribution with zero variance is a point mass on its mean.

- There exists a random ℓ-vector $\mathbf{z}$, whose components are independent standard normal random variables, a k-vector $\boldsymbol{\mu}$, and a $k \times \ell$ matrix $\mathbf{A}$, such that $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$. Here ℓ is the rank of the covariance matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$. Especially in the case of full rank, see the section below on Geometric interpretation.
- There is a k-vector $\boldsymbol{\mu}$ and a symmetric, nonnegative-definite $k \times k$ matrix $\boldsymbol{\Sigma}$, such that the characteristic function of $\mathbf{x}$ is

$$\varphi_{\mathbf{x}}(\mathbf{u}) = \exp\left(i\mathbf{u}'\boldsymbol{\mu} - \tfrac{1}{2}\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}\right).$$

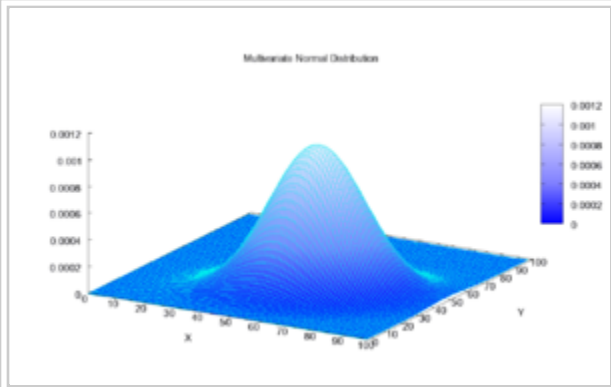
The covariance matrix is allowed to be singular (in which case the corresponding distribution has no density). This case arises frequently in statistics; for example, in the distribution of the vector of residuals in the ordinary least squares regression. Note also that the X_i are in general *not* independent; they can be seen as the result of applying the matrix $\mathbf{A}$ to a collection of independent Gaussian variables $\mathbf{z}$.

Properties

Density function

Non-degenerate case

The multivariate normal distribution is said to be "non-degenerate" when the symmetric covariance matrix $\boldsymbol{\Sigma}$ is positive definite. In this case the distribution has density[2]


Bivariate normal joint density

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

where $\mathbf{x}$ is a real k-dimensional column vector and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. Note how the equation above reduces to that of the univariate normal distribution if $\boldsymbol{\Sigma}$ is a 1×1 matrix (i.e. a single real number).

Note that the circularly-symmetric version of the complex normal distribution has a slightly different form.

Each iso-density locus—the locus of points in k-dimensional space each of which gives the same

particular value of the density—is an ellipse or its higher-dimensional generalization; hence the multivariate normal is a special case of the elliptical distributions.

The descriptive statistic $(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ in the non-degenerate multivariate normal distribution equation is known as the square of the Mahalanobis distance, which represents the distance of the test point $\mathbf{x}$ from the mean $\boldsymbol{\mu}$. Note that in case when $k = 1$, the distribution reduces to a univariate normal distribution and the Mahalanobis distance reduces to the standard score.

Bivariate case

In the 2-dimensional nonsingular case (k = rank(Σ) = 2), the probability density function of a vector [X Y]′ is:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{2\rho(x-}\right.\right.$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{x-\mu_X}{\sigma_X}-\frac{y-\mu_Y}{\sigma_Y}\right]^2\right),$$

where ϱ is the correlation between X and Y and where $\sigma_X > 0$ and $\sigma_Y > 0$. In this case,

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}.$$

In the bivariate case, the first equivalent condition for multivariate normality can be made less restrictive: it is sufficient to verify that countably many distinct linear combinations of X and Y are normal in order to conclude that the vector [X Y]′ is bivariate normal.[3]

The bivariate iso-density loci plotted in the x,y-plane are ellipses. As the correlation parameter ϱ increases, these loci appear to be squeezed to the following line :

$$y(x) = \operatorname{sgn}(\rho)\frac{\sigma_Y}{\sigma_X}(x-\mu_X)+\mu_Y.$$

This is because the above expression - but without ϱ being inside a signum function - is the best linear unbiased prediction of Y given a value of X.[4]

Degenerate case

If the covariance matrix $\boldsymbol{\Sigma}$ is not full rank, then the multivariate normal distribution is degenerate and does not have a density. More precisely, it does not have a density with respect to k-dimensional Lebesgue measure (which is the usual measure assumed in calculus-level probability courses). Only random vectors whose distributions are absolutely continuous with respect to a measure are said to have densities (with respect to that measure). To talk about densities but avoid dealing with measure-theoretic complications it can be simpler to restrict attention to a subset of $\operatorname{rank}(\boldsymbol{\Sigma})$ of the coordinates of $\mathbf{x}$ such that the covariance matrix for this subset is positive definite; then the other coordinates may be thought

of as an affine function of the selected coordinates.

To talk about densities meaningfully in the singular case, then, we must select a different base measure. Using the disintegration theorem we can define a restriction of Lebesgue measure to the $\operatorname{rank}(\boldsymbol\Sigma)$ -dimensional affine subspace of $\mathbb{R}^k$ where the Gaussian distribution is supported, i.e. $\{\boldsymbol\mu + \boldsymbol\Sigma^{1/2}\mathbf{v} : \mathbf{v} \in \mathbb{R}^k\}$. With respect to this measure the distribution has density:

$$f(\mathbf{x}) = \left(\overset{*}{\det}(2\pi\boldsymbol\Sigma)\right)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol\mu)'\boldsymbol\Sigma^{+}(\mathbf{x}-\boldsymbol\mu)}$$

where $\boldsymbol\Sigma^{+}$ is the generalized inverse and det* is the pseudo-determinant.[5]

Higher moments

The *k*th-order moments of **x** are defined by

$$\mu_{1,\ldots,N}(\mathbf{x}) \overset{\text{def}}{=} \mu_{r_1,\ldots,r_N}(\mathbf{x}) \overset{\text{def}}{=} E\left[\prod_{j=1}^{N} x_j^{r_j}\right]$$

where $r_1 + r_2 + \cdots + r_N = k$.

The central *k*-order central moments are given as follows

(a) If *k* is odd, $\mu_{1,\ldots,N}(\mathbf{x} - \boldsymbol\mu) = 0$.

(b) If *k* is even with $k = 2\lambda$, then

$$\mu_{1,\ldots,2\lambda}(\mathbf{x} - \boldsymbol\mu) = \sum (\Sigma_{ij}\Sigma_{k\ell}\cdots\Sigma_{XZ})$$

where the sum is taken over all allocations of the set $\{1, \ldots, 2\lambda\}$ into λ (unordered) pairs. That is, if you have a *k*th ($= 2\lambda = 6$) central moment, you will be summing the products of $\lambda = 3$ covariances (the -**μ** notation has been dropped in the interests of parsimony):

$$\begin{aligned}
&E[x_1x_2x_3x_4x_5x_6] \\
&= E[x_1x_2]E[x_3x_4]E[x_5x_6] + E[x_1x_2]E[x_3x_5]E[x_4x_6] + E[x_1x_2]E[x_3x_6]E[x_4 \\
&+ E[x_1x_3]E[x_2x_4]E[x_5x_6] + E[x_1x_3]E[x_2x_5]E[x_4x_6] + E[x_1x_3]E[x_2x_6]E[x_4 \\
&+ E[x_1x_4]E[x_2x_3]E[x_5x_6] + E[x_1x_4]E[x_2x_5]E[x_3x_6] + E[x_1x_4]E[x_2x_6]E[x_3 \\
&+ E[x_1x_5]E[x_2x_3]E[x_4x_6] + E[x_1x_5]E[x_2x_4]E[x_3x_6] + E[x_1x_5]E[x_2x_6]E[x_3 \\
&+ E[x_1x_6]E[x_2x_3]E[x_4x_5] + E[x_1x_6]E[x_2x_4]E[x_3x_5] + E[x_1x_6]E[x_2x_5]E[x_3
\end{aligned}$$

This yields $(2\lambda - 1)!/(2^{\lambda-1}(\lambda - 1)!)$ terms in the sum (15 in the above case), each being the

product of λ (in this case 3) covariances. For fourth order moments (four variables) there are three terms. For sixth-order moments there are $3 \times 5 = 15$ terms, and for eighth-order moments there are $3 \times 5 \times 7 = 105$ terms.

The covariances are then determined by replacing the terms of the list $[1, \ldots, 2\lambda]$ by the corresponding terms of the list consisting of r_1 ones, then r_2 twos, etc.. To illustrate this, examine the following 4th-order central moment case:

$$\begin{aligned}
E\left[x_i^4\right] &= 3\Sigma_{ii}^2 \\
E\left[x_i^3 x_j\right] &= 3\Sigma_{ii}\Sigma_{ij} \\
E\left[x_i^2 x_j^2\right] &= \Sigma_{ii}\Sigma_{jj} + 2\left(\Sigma_{ij}\right)^2 \\
E\left[x_i^2 x_j x_k\right] &= \Sigma_{ii}\Sigma_{jk} + 2\Sigma_{ij}\Sigma_{ik} \\
E\left[x_i x_j x_k x_n\right] &= \Sigma_{ij}\Sigma_{kn} + \Sigma_{ik}\Sigma_{jn} + \Sigma_{in}\Sigma_{jk}.
\end{aligned}$$

where Σ_{ij} is the covariance of x_i and x_j. The idea with the above method is you first find the general case for a kth moment where you have k different x variables - $E\left[x_i x_j x_k x_n\right]$ and then you can simplify this accordingly. Say, you have $E\left[x_i^2 x_k x_n\right]$ then you simply let $x_i = x_j$ and realise that $\Sigma_{ii} = \sigma_i^2$.

Likelihood function

If the mean and variance matrix are unknown, a suitable log likelihood function for a single observation **x** would be:

$$\ln(L) = -\frac{1}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{k}{2}\ln(2\pi)$$

where x is a vector of real numbers. The circularly-symmetric version of the complex case, where z is a vector of complex numbers, would be

$$\ln(L) = -\ln(|\boldsymbol{\Sigma}|) - (\mathbf{z}-\boldsymbol{\mu})^{\dagger}\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) - k\ln(\pi)$$

i.e. with the conjugate transpose (indicated by †) replacing the normal transpose (indicated by T). This is slightly different than in the real case, because the circularly-symmetric version of the complex normal distribution has a slightly different form.

A similar notation is used for multiple linear regression.[6]

Entropy

The differential entropy of the multivariate normal distribution is[7]

$$h(f) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\ln f(\mathbf{x})\,d\mathbf{x},$$
$$= \frac{1}{2}\ln\left((2\pi e)^n\cdot|\boldsymbol{\Sigma}|\right),$$

where the bars denote the matrix determinant.

Kullback–Leibler divergence

The Kullback–Leibler divergence from $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ to $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, for non-singular matrices Σ_0 and Σ_1, is:[8]

$$D_{\mathrm{KL}}(\mathcal{N}_0\|\mathcal{N}_1) = \frac{1}{2}\left\{\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) - K + \ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right.$$

where K is the dimension of the vector space.

The logarithm must be taken to base *e* since the two terms following the logarithm are themselves base-*e* logarithms of expressions that are either factors of the density function or otherwise arise naturally. The equation therefore gives a result measured in nats. Dividing the entire expression above by $\log_e 2$ yields the divergence in bits.

Cumulative distribution function

The notion of cumulative distribution function (cdf) in dimension 1 can be extended in two ways to the multidimensional case. The first way is to define the cumulative distribution function $F(r)$ as the probability that a sample **falls** inside the ellipsoid determined by its Mahalanobis distance r from the Gaussian, a direct generalization of the standard deviation .[9] In order to compute the values of this function, closed analytic formulae exist.[9]

Another way to extend the notion of cumulative distribution function is to define the cumulative distribution function (cdf) $F(\mathbf{x}_0)$ of a random vector $\mathbf{x}$ as the probability that all components of $\mathbf{x}$ are less than or equal to the corresponding values in the vector $\mathbf{x}_0$. Though there is no closed form for $F(\mathbf{x})$, there are a number of algorithms that estimate it numerically.[10]

Prediction Interval

The prediction interval for the multivariate normal distribution yields a region consisting of those vectors $\mathbf{x}$ satisfying

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \le \chi_k^2(p).$$

Here $\mathbf{x}$ is a k-dimensional vector, $\boldsymbol{\mu}$ is the known k-dimensional mean vector, $\boldsymbol{\Sigma}$ is the known covariance matrix and $\chi_k^2(p)$ is the quantile function for probability p of the chi-squared distribution

with k degrees of freedom.[11]

When $k = 2,$ the expression defines the interior of an ellipse and the chi-squared distribution simplifies to an exponential distribution with mean equal to two.

Joint normality

Normally distributed and independent

If *X* and *Y* are normally distributed and independent, this implies they are "jointly normally distributed", i.e., the pair (*X*, *Y*) must have multivariate normal distribution. However, a pair of jointly normally distributed variables need not be independent (would only be so if uncorrelated, $\rho = 0$).

Two normally distributed random variables need not be jointly bivariate normal

The fact that two random variables *X* and *Y* both have a normal distribution does not imply that the pair (*X*, *Y*) has a joint normal distribution. A simple example is one in which X has a normal distribution with expected value 0 and variance 1, and $Y = X$ if $|X| > c$ and $Y = -X$ if $|X| < c$, where $c > 0$. There are similar counterexamples for more than two random variables. In general, they sum to a mixture model.

Correlations and independence

In general, random variables may be uncorrelated but statistically dependent. But if a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent. This implies that any two or more of its components that are pairwise independent are independent.

But it is **not** true that two random variables that are (separately, marginally) normally distributed and uncorrelated are independent. Two random variables that are normally distributed may fail to be *jointly* normally distributed, i.e., the vector whose components they are may fail to have a multivariate normal distribution. In the preceding example, clearly *X* and *Y* are not independent, yet choosing *c* to be 1.54 makes them uncorrelated.

Conditional distributions

If N-dimensional **x** is partitioned as follows

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

and accordingly **μ** and **Σ** are partitioned as follows

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

then, the distribution of $\mathbf{x}_1$ conditional on $\mathbf{x}_2 = a$ is multivariate normal $(\mathbf{x}_1|\mathbf{x}_2 = \mathbf{a}) \sim N(\overline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ where

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{a} - \boldsymbol{\mu}_2\right)$$

and covariance matrix

$$\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$[12]

This matrix is the Schur complement of $\boldsymbol{\Sigma}_{22}$ in $\boldsymbol{\Sigma}$. This means that to calculate the conditional covariance matrix, one inverts the overall covariance matrix, drops the rows and columns corresponding to the variables being conditioned upon, and then inverts back to get the conditional covariance matrix. Here $\boldsymbol{\Sigma}_{22}^{-1}$ is the generalized inverse of $\boldsymbol{\Sigma}_{22}$.

Note that knowing that $\mathbf{x}_2 = \mathbf{a}$ alters the variance, though the new variance does not depend on the specific value of $\mathbf{a}$; perhaps more surprisingly, the mean is shifted by $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{a} - \boldsymbol{\mu}_2\right)$; compare this with the situation of not knowing the value of $\mathbf{a}$, in which case $\mathbf{x}_1$ would have distribution $\mathcal{N}_q\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}\right)$.

An interesting fact derived in order to prove this result, is that the random vectors $\mathbf{x_2}$ and $\mathbf{y}_1 = \mathbf{x}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2$ are independent.

The matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}{}^{-1}$ is known as the matrix of regression coefficients.

Bivariate case

In the bivariate case where $\mathbf{x}$ is partitioned into X_1 and X_2, the conditional distribution of X_1 given X_2 is[13]

$$X_1|X_2 = x_2 \ \sim \ \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2),\ (1-\rho^2)\sigma_1^2\right).$$

where ρ is the correlation coefficient between X_1 and X_2.

Bivariate conditional expectation

In the general case

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

The conditional expectation of X_1 given X_2 is:

$$\mathrm{E}(X_1|X_2 = x_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$$

Proof: the result is simply obtained taking the expectation of the conditional distribution $X_1|X_2$ above.

In the case of unit variances

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

The conditional expectation of X_1 given X_2 is:

$$\mathrm{E}(X_1 \mid X_2 = x_2) = \rho x_2$$

and the conditional expectation of X_1 given that X_2 is smaller/bigger than z is (Maddala 1983, p. 367[14]) :

$$\mathrm{E}(X_1 \mid X_2 < z) = -\rho\frac{\phi(z)}{\Phi(z)},$$

$$\mathrm{E}(X_1 \mid X_2 > z) = \rho\frac{\phi(z)}{(1 - \Phi(z))},$$

where the final ratio here is called the inverse Mills ratio.

Proof: the last two results are obtained using the result $\mathrm{E}(X_1 \mid X_2 = x_2) = \rho x_2$, so that

$\mathrm{E}(X_1 \mid X_2 < z) = \rho E(X_2 \mid X_2 < z)$ and then using the properties of the expectation of a truncated normal distribution.

Marginal distributions

To obtain the marginal distribution over a subset of multivariate normal random variables, one only needs to drop the irrelevant variables (the variables that one wants to marginalize out) from the mean vector and the covariance matrix. The proof for this follows from the definitions of multivariate normal distributions and linear algebra.[15]

Example

Let $\mathbf{x} = [X_1, X_2, X_3]$ be multivariate normal random variables with mean vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3]$ and covariance matrix $\boldsymbol{\Sigma}$ (standard parametrization for multivariate normal distributions). Then the joint distribution of $\mathbf{x}' = [X_1, X_3]$ is multivariate normal with mean vector $\boldsymbol{\mu}' = [\mu_1, \mu_3]$ and covariance matrix

$$\boldsymbol{\Sigma}' = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{13} \\ \boldsymbol{\Sigma}_{31} & \boldsymbol{\Sigma}_{33} \end{bmatrix}.$$

Affine transformation

If **y** = **c** + **Bx** is an affine transformation of $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where **c** is an $M \times 1$ vector of constants and **B** is a constant $M \times N$ matrix, then **y** has a multivariate normal distribution with expected value **c** + **Bμ** and variance **BΣB**$^{\mathrm{T}}$ i.e., $\mathbf{y} \sim \mathcal{N}\left(\mathbf{c} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\mathrm{T}}\right)$. In particular, any subset of the x_i has a marginal distribution that is also multivariate normal. To see this, consider the following example: to extract the subset $(x_1, x_2, x_4)^{\mathrm{T}}$, use

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$$

which extracts the desired elements directly.

Another corollary is that the distribution of **Z** = **b** · **x**, where **b** is a constant vector of the same length as **x** and the dot indicates a vector product, is univariate Gaussian with $Z \sim \mathcal{N}\left(\mathbf{b} \cdot \boldsymbol{\mu}, \mathbf{b}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{b}\right)$. This result follows by using

$$\mathbf{B} = \begin{bmatrix} b_1 & b_2 & \dots & b_n \end{bmatrix} = \mathbf{b}^{\mathrm{T}}.$$

Observe how the positive-definiteness of **Σ** implies that the variance of the dot product must be positive.

An affine transformation of **x** such as 2**x** is not the same as the sum of two independent realisations of **x**.

Geometric interpretation

The equidensity contours of a non-singular multivariate normal distribution are ellipsoids (i.e. linear transformations of hyperspheres) centered at the mean.[16] Hence the multivariate normal distribution is an example of the class of elliptical distributions. The directions of the principal axes of the ellipsoids are given by the eigenvectors of the covariance matrix **Σ**. The squared relative lengths of the principal axes are given by the corresponding eigenvalues.

If **Σ** = **UΛU**$^{\mathrm{T}}$ = **UΛ**$^{1/2}$(**UΛ**$^{1/2}$)$^{\mathrm{T}}$ is an eigendecomposition where the columns of **U** are unit eigenvectors and **Λ** is a diagonal matrix of the eigenvalues, then we have

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff \mathbf{x} \sim \boldsymbol{\mu} + \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathcal{N}(0, \mathbf{I}) \iff \mathbf{x} \sim \boldsymbol{\mu} + \mathbf{U}\mathcal{N}(0, \boldsymbol{\Lambda}$$

Moreover, **U** can be chosen to be a rotation matrix, as inverting an axis does not have any effect on $N(0, \boldsymbol{\Lambda})$, but inverting a column changes the sign of **U**'s determinant. The distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is in effect $N(0, \mathbf{I})$ scaled by $\boldsymbol{\Lambda}^{1/2}$, rotated by **U** and translated by **μ**.

Conversely, any choice of $\boldsymbol{\mu}$, full rank matrix $\mathbf{U}$, and positive diagonal entries Λ_i yields a non-singular multivariate normal distribution. If any Λ_i is zero and $\mathbf{U}$ is square, the resulting covariance matrix $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$ is singular. Geometrically this means that every contour ellipsoid is infinitely thin and has zero volume in n-dimensional space, as at least one of the principal axes has length of zero.

Estimation of parameters

The derivation of the maximum-likelihood estimator of the covariance matrix of a multivariate normal distribution is perhaps surprisingly subtle and elegant. See estimation of covariance matrices.

In short, the probability density function (pdf) of a multivariate normal is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

and the ML estimator of the covariance matrix from a sample of n observations is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

which is simply the sample covariance matrix. This is a biased estimator whose expectation is

$$E[\widehat{\boldsymbol{\Sigma}}] = \frac{n-1}{n}\boldsymbol{\Sigma}.$$

An unbiased sample covariance is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}.$$

The Fisher information matrix for estimating the parameters of a multivariate normal distribution has a closed form expression. This can be used, for example, to compute the Cramér–Rao bound for parameter estimation in this setting. See Fisher information for more details.

Bayesian inference

In Bayesian statistics, the conjugate prior of the mean vector is another multivariate normal distribution, and the conjugate prior of the covariance matrix is an inverse-Wishart distribution $\mathcal{W}^{-1}$. Suppose then that n observations have been made

$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and that a conjugate prior has been assigned, where

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma})\ p(\boldsymbol{\Sigma}),$$

where

$$p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) \sim \mathcal{N}(\boldsymbol{\mu}_0, m^{-1}\boldsymbol{\Sigma}),$$

and

$$p(\boldsymbol{\Sigma}) \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, n_0).$$

Then,

$$\begin{aligned} p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \mathbf{X}) &\sim \mathcal{N}\left(\tfrac{n\bar{\mathbf{x}}+m\boldsymbol{\mu}_0}{n+m}, \tfrac{1}{n+m}\boldsymbol{\Sigma}\right), \\ p(\boldsymbol{\Sigma} \mid \mathbf{X}) &\sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi} + n\mathbf{S} + \tfrac{nm}{n+m}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)', n+n_0\right), \end{aligned}$$

where

$$\begin{aligned} \bar{\mathbf{x}} &= n^{-1}\textstyle\sum_{i=1}^{n}\mathbf{x}_i, \\ \mathbf{S} &= n^{-1}\textstyle\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'. \end{aligned}$$

Multivariate normality tests

Multivariate normality tests check a given set of data for similarity to the multivariate normal distribution. The null hypothesis is that the data set is similar to the normal distribution, therefore a sufficiently small *p*-value indicates non-normal data. Multivariate normality tests include the Cox-Small test[17] and Smith and Jain's adaptation[18] of the Friedman-Rafsky test.[19]

Mardia's test[20] is based on multivariate extensions of skewness and kurtosis measures. For a sample $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ of *k*-dimensional vectors we compute

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})^T$$

$$A = \frac{1}{6n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\mathbf{x}_i-\bar{\mathbf{x}})^T\,\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_j-\bar{\mathbf{x}})\right]^3$$

$$B = \sqrt{\frac{n}{8k(k+2)}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{x}_i-\bar{\mathbf{x}})^T\,\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\right]^2 - k(k+2)\right\}$$

Under the null hypothesis of multivariate normality, the statistic *A* will have approximately a chi-squared distribution with $\frac{1}{6}\cdot k(k+1)(k+2)$ degrees of freedom, and *B* will be approximately standard normal *N*(0,1).

Mardia's kurtosis statistic is skewed and converges very slowly to the limiting normal distribution. For medium size samples $(50 \leq n < 400)$, the parameters of the asymptotic distribution of the kurtosis statistic are modified[21] For small sample tests ($n < 50$) empirical critical values are used. Tables of critical values for both statistics are given by Rencher[22] for *k*=2,3,4.

Mardia's tests are affine invariant but not consistent. For example, the multivariate skewness test is not consistent against symmetric non-normal alternatives.[23]

The **BHEP test**[24] computes the norm of the difference between the empirical characteristic function and the theoretical characteristic function of the normal distribution. Calculation of the norm is performed in the $L^2(\mu)$ space of square-integrable functions with respect to the Gaussian weighting function $\mu_\beta(\mathbf{t}) = (2\pi\beta^2)^{-k/2} e^{-|\mathbf{t}|^2/(2\beta^2)}$. The test statistic is

$$T_\beta = \int_{\mathbb{R}^k} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\mathbf{t}^T \widehat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}})} - e^{-|\mathbf{t}|^2/2} \right|^2 \boldsymbol{\mu}_\beta(\mathbf{t}) d\mathbf{t}$$

$$= \frac{1}{n^2} \sum_{i,j=1}^{n} e^{-\frac{\beta^2}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \mathbf{x}_j)} - \frac{2}{n(1+\beta^2)^{k/2}} \sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}(\mathbf{x}_i - \bar{\mathbf{x}})^T \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i -$$

The limiting distribution of this test statistic is a weighted sum of chi-squared random variables,[24] however in practice it is more convenient to compute the sample quantiles using the Monte-Carlo simulations.

A detailed survey of these and other test procedures is available.[25]

Drawing values from the distribution

A widely used method for drawing (sampling) a random vector **x** from the *N*-dimensional multivariate normal distribution with mean vector **μ** and covariance matrix **Σ** works as follows:[26]

1. Find any real matrix **A** such that $\mathbf{A}\,\mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma}$. When **Σ** is positive-definite, the Cholesky decomposition is typically used, and the extended form of this decomposition can always be used (as the covariance matrix may be only positive semi-definite) in both cases a suitable matrix **A** is obtained. An alternative is to use the matrix $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}$ obtained from a spectral decomposition $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$ of **Σ**. The former approach is more computationally straightforward but the matrices **A** change for different orderings of the elements of the random vector, while the latter approach gives matrices that are related by simple re-orderings. In theory both approaches give equally good ways of determining a suitable matrix **A**, but there are differences in computation time.
2. Let $\mathbf{z} = (z_1, \ldots, z_N)^{\mathrm{T}}$ be a vector whose components are *N* independent standard normal variates (which can be generated, for example, by using the Box–Muller transform).
3. Let **x** be **μ** + **Az**. This has the desired distribution due to the affine transformation property.

See also

- Chi distribution, the pdf of the 2-norm (or Euclidean norm) of a multivariate normally distributed vector (centered at zero).
- Complex normal distribution, for the generalization to complex valued random variables.
- Copula, for the definition of the Gaussian or normal copula model.
- Multivariate stable distribution extension of the multivariate normal distribution, when the index (exponent in the characteristic function) is between zero to two.
- Mahalanobis distance
- Wishart distribution

References

1. Gut, Allan (2009) *An Intermediate Course in Probability*, Springer. ISBN 9781441901613 (Chapter 5)
2. UIUC, Lecture 21. *The Multivariate Normal Distribution* (http://www.math.uiuc.edu/~r-ash/Stat/StatLec21-25.pdf), 21.5:"Finding the Density".
3. Hamedani, G. G.; Tata, M. N. (1975). "On the determination of the bivariate normal distribution from distributions of linear combinations of the variables". *The American Mathematical Monthly* **82** (9): 913–915. doi:10.2307/2318494 (https://dx.doi.org/10.2307%2F2318494).
4. Wyatt, John. "Linear least mean-squared error estimation" (http://web.mit.edu/6.041/www/LECTURE/lec22.pdf) (PDF). *Lecture notes course on applied probability*. Retrieved 23 January 2012.
5. Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley. pp. 527–528.
6. Tong, T. (2010) Multiple Linear Regression : MLE and Its Distributional Results (http://amath.colorado.edu/courses/7400/2010Spr/lecture9.pdf), Lecture Notes
7. Gokhale, DV; Ahmed, NA; Res, BC; Piscataway, NJ (May 1989). "Entropy Expressions and Their Estimators for Multivariate Distributions". *Information Theory, IEEE Transactions on* **35** (3): 688–692. doi:10.1109/18.30996 (https://dx.doi.org/10.1109%2F18.30996).
8. J. Duchi, Derivations for Linear Algebra and Optimization [1] (http://www.cs.berkeley.edu/~jduchi/projects/general_notes.pdf). pp. 13
9. Bensimhoun Michael, *N-Dimensional Cumulative Function, And Other Useful Facts About Gaussians and Normal Densities* (2006) (https://upload.wikimedia.org/wikipedia/commons/a/a2/Cumulative_function_n_dimensional_Gaussians_12.2013.pdf)
10. Genz, Alan (2009). *Computation of Multivariate Normal and t Probabilities* (http://www.springer.com/statistics/computational+statistics/book/978-3-642-01688-2). Springer. ISBN 978-3-642-01689-9.
11. Siotani, Minoru (1964). "Tolerance regions for a multivariate normal population" (http://www.ism.ac.jp/editsec/aism/pdf/016_1_0135.pdf) (PDF). *Annals of the Institute of Statistical Mathematics* **16** (1): 135–153. doi:10.1007/BF02868568 (https://dx.doi.org/10.1007%2FBF02868568).
12. Eaton, Morris L. (1983). *Multivariate Statistics: a Vector Space Approach*. John Wiley and Sons. pp. 116–117. ISBN 0-471-02776-6.
13. Jensen, J (2000). *Statistics for Petroleum Engineers and Geoscientists*. Amsterdam: Elsevier. p. 207.
14. Gangadharrao, Maddala (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
15. The formal proof for marginal distribution is shown here http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html
16. Nikolaus Hansen. "The CMA Evolution Strategy: A Tutorial" (http://www.lri.fr/~hansen/cmatutorial.pdf) (PDF).
17. Cox, D. R.; Small, N. J. H. (1978). "Testing multivariate normality". *Biometrika* **65** (2): 263. doi:10.1093/biomet/65.2.263 (https://dx.doi.org/10.1093%2Fbiomet%2F65.2.263).
18. Smith, S. P.; Jain, A. K. (1988). "A test to determine the multivariate normality of a data set". *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10** (5): 757. doi:10.1109/34.6789 (https://dx.doi.org/10.1109%2F34.6789).

19. Friedman, J. H.; Rafsky, L. C. (1979). "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests". *The Annals of Statistics* **7** (4): 697. doi:10.1214/aos/1176344722 (https://dx.doi.org/10.1214%2Faos%2F1176344722).
20. Mardia, K. V. (1970). "Measures of multivariate skewness and kurtosis with applications". *Biometrika* **57** (3): 519–530. doi:10.1093/biomet/57.3.519 (https://dx.doi.org/10.1093%2Fbiomet%2F57.3.519).
21. Rencher (1995), pages 112-113.
22. Rencher (1995), pages 493-495.
23. Baringhaus, L.; Henze, N. (1991). "Limit distributions for measures of multivariate skewness and kurtosis based on projections". *Journal of Multivariate Analysis* **38**: 51. doi:10.1016/0047-259X(91)90031-V (https://dx.doi.org/10.1016%2F0047-259X%2891%2990031-V).
24. Baringhaus, L.; Henze, N. (1988). "A consistent test for multivariate normality based on the empirical characteristic function". *Metrika* **35** (1): 339–348. doi:10.1007/BF02613322 (https://dx.doi.org/10.1007%2FBF02613322).
25. Henze, Norbert (2002). "Invariant tests for multivariate normality: a critical review". *Statistical Papers* **43** (4): 467–506. doi:10.1007/s00362-002-0119-6 (https://dx.doi.org/10.1007%2Fs00362-002-0119-6).
26. Gentle, J.E. (2009). *Computational Statistics*. New York: Springer. pp. 315–316. doi:10.1007/978-0-387-98144-4 (https://dx.doi.org/10.1007%2F978-0-387-98144-4).

Literature

- Rencher, A.C. (1995). *Methods of Multivariate Analysis*. New York: Wiley.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Multivariate_normal_distribution&oldid=672552874"

Categories: Continuous distributions | Multivariate continuous distributions | Normal distribution | Exponential family distributions | Stable distributions | Probability distributions

---

- This page was last modified on 22 July 2015, at 09:03.
- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.