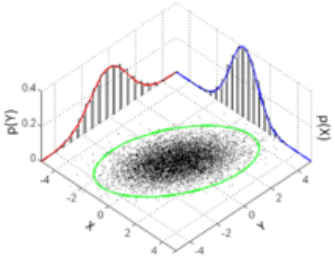
Multivariate normal distribution

In probability theory and statistics, the **multivariate normal distribution**, **multivariate Gaussian distribution**, or **joint normal distribution** is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. One 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

Multivariate normal

Probability density function



Many sample points from a multivariate normal distribution with $\boldsymbol{\mu} = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$ and $\boldsymbol{\Sigma} = \left[\begin{smallmatrix} 1 & 3/5 \\ 3/5 & 2 \end{smallmatrix}\right]$, 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 (positive semi-definite matrix)
Support	$\mathbf{x} \in \boldsymbol{\mu} + \operatorname{span}(\boldsymbol{\Sigma}) \subseteq \mathbf{R}^k$
PDF	$\det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}}\ \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$, exists only when $\boldsymbol{\Sigma}$ is positive-definite
Mean	$\boldsymbol{\mu}$
Mode	$\boldsymbol{\mu}$
Variance	$\boldsymbol{\Sigma}$
Entropy	$\frac{1}{2}\ln\det\left(2\pi\mathrm{e}\boldsymbol{\Sigma}\right)$
MGF	$\exp\left(\boldsymbol{\mu}^{\mathrm{T}}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t}\right)$
CF	$\exp\left(i\boldsymbol{\mu}^{\mathrm{T}}\mathbf{t} - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{t}\right)$
Kullback-Leibler divergence	see below

Definitions

Notation and parameterization

The multivariate normal distribution of a *k*-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)^{\mathbf{T}}$ 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

$$\boldsymbol{\mu} = \mathbf{E}[\mathbf{X}] = (\mathbf{E}[X_1], \mathbf{E}[X_2], \ldots, \mathbf{E}[X_k])^{\mathbf{T}},$$

and $k \times k$ covariance matrix

$$\boldsymbol{\Sigma}_{i,j} = \mathbf{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j]$$

such that $1 \leq i, j \leq k$. The inverse of the covariance matrix is called the precision matrix, denoted by $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$.

Standard normal random vector

A real random vector $\mathbf{X} = (X_1, \ldots, X_k)^{\mathbf{T}}$ is called a **standard normal random vector** if all of its components X_k are independent and each is a zero-mean unit-variance normally distributed random variable, i.e. if $X_k \sim \mathcal{N}(0, 1)$ for all k.[1]:p. 454

Centered normal random vector

A real random vector $\mathbf{X} = (X_1, \ldots, X_k)^{\mathbf{T}}$ is called a **centered normal random vector** if there exists a deterministic $k \times \ell$ matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\mathbf{Z}$ has the same distribution as $\mathbf{X}$ where $\mathbf{Z}$ is a standard normal random vector with ℓ components.[1]:p. 454

Normal random vector

A real random vector $\mathbf{X} = (X_1, \ldots, X_k)^{\mathbf{T}}$ is called a **normal random vector** if there exists a random ℓ-vector $\mathbf{Z}$, which is a standard normal random vector, a k-vector $\boldsymbol{\mu}$, and a $k \times \ell$ matrix $\boldsymbol{A}$, such that $\mathbf{X} = \boldsymbol{A}\mathbf{Z} + \boldsymbol{\mu}$.[2]:p. 454[1]:p. 455

Formally:

$$\mathbf{X} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma}) \iff \text{there exist } \mu \in \mathbb{R}^k, \boldsymbol{A} \in \mathbb{R}^{k \times \ell} \text{ such that } \mathbf{X} = \boldsymbol{A}\mathbf{Z} + \mu \text{ and } \forall n = 1, \ldots, l : Z_n \sim \mathcal{N}(0, 1), \text{i.i.d.}$$

Here the covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^{\mathbf{T}}$.

In the degenerate case where the covariance matrix is singular, the corresponding distribution has no density; see the section below for details. This case arises frequently in statistics; for example, in the distribution of the vector of residuals in the ordinary least squares regression. The $\boldsymbol{X_i}$ are in general *not* independent; they can be seen as the result of applying the matrix $\boldsymbol{A}$ to a collection of independent Gaussian variables $\mathbf{Z}$.

Equivalent definitions

The following definitions are equivalent to the definition given above. A random vector $\mathbf{X} = (X_1, \ldots, X_k)^T$ has a multivariate normal distribution if it satisfies one of the following equivalent conditions.

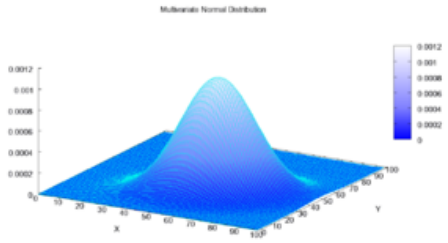
- Every linear combination $Y = a_1 X_1 + \cdots + a_k X_k$ of its components is normally distributed. That is, for any constant vector $\mathbf{a} \in \mathbb{R}^k$, the random variable $Y = \mathbf{a}^{\mathrm{T}}\mathbf{X}$ has a univariate normal distribution, where a univariate normal distribution with zero variance is a point mass on its mean.
- There is a *k*-vector $\boldsymbol{\mu}$ and a symmetric, positive semidefinite $k \times k$ matrix $\boldsymbol{\Sigma}$, such that the characteristic function of $\mathbf{X}$ is

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

The spherical normal distribution can be characterised as the unique distribution where components are independent in any orthogonal coordinate system.[3][4]

Density function

Non-degenerate case



Bivariate normal joint density

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[5]

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

where $\mathbf{x}$ is a real *k*-dimensional column vector and $|\boldsymbol{\Sigma}| \equiv \det \boldsymbol{\Sigma}$ is the determinant of $\boldsymbol{\Sigma}$, also known as the generalized variance. 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).

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 quantity $\sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$ is known as the Mahalanobis distance, which represents the distance of the test point $\mathbf{x}$ from the mean $\boldsymbol{\mu}$. Note that in the case when $k = 1$, the distribution reduces to a univariate normal distribution and the Mahalanobis distance reduces to the absolute value of the standard score. See also Interval below.

Bivariate case

In the 2-dimensional nonsingular case ($k = \operatorname{rank}(\Sigma) = 2$), the probability density function of a vector $[\mathbf{XY}]'$ is:

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

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

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

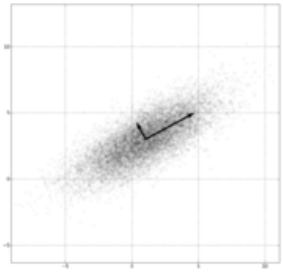
In the bivariate case, the first equivalent condition for multivariate reconstruction of normality can be made less restrictive as it is sufficient to verify that countably many distinct linear combinations of X and Y are normal in order to conclude that the vector of $[\mathbf{XY}]'$ is bivariate normal.[6]

The bivariate iso-density loci plotted in the x, y-plane are ellipses, whose principal axes are defined by the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ (the major and minor semidiameters of the ellipse equal the square-root of the ordered eigenvalues).

As the absolute value of the correlation parameter ρ increases, these loci are squeezed toward the following line :

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

This is because this expression, with $\operatorname{sgn}(\rho)$ (where sgn is the Sign function) replaced by ρ, is the best linear unbiased prediction of Y given a value of X.[7]



Bivariate normal distribution centered at $(1, 3)$ with a standard deviation of 3 in roughly the $(0.878, 0.478)$ direction and of 1 in the orthogonal direction.

Degenerate case

If the covariance matrix Σ 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 $\operatorname{rank}(\Sigma)$ 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 these selected coordinates.

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

$$f(\mathbf{x}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{+}(\mathbf{x}-\mu)}}{\sqrt{(2\pi)^k \det{}^*(\Sigma)}}$$

where Σ^{+} is the generalized inverse, k is the rank of Σ and det^* is the pseudo-determinant.[8]

Cumulative distribution function

The notion of cumulative distribution function (cdf) in dimension 1 can be extended in two ways to the multidimensional case, based on rectangular and ellipsoidal regions.

The first way is to define the cdf $F(\mathbf{x})$ 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 $\mathbf{x}$:[9]

$$F(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}), \quad \text{where } \mathbf{X} \sim \mathcal{N}(\mu, \Sigma).$$

Though there is no closed form for $F(\mathbf{x})$, there are a number of algorithms that estimate it numerically (https://cran.r-project.org/web/packages/TruncatedNormal/).[9][10]

Another way is to define the cdf $F(r)$ as the probability that a sample lies inside the ellipsoid determined by its Mahalanobis distance r from the Gaussian, a direct generalization of the standard deviation.[11] In order to compute the values of this function, closed analytic formulae exist,[11] as follows.

Interval

The interval for the multivariate normal distribution yields a region consisting of those vectors **x** satisfying

$$(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \leq \chi^2_k(p).$$

Here $\mathbf{x}$ is a k-dimensional vector, μ is the known k-dimensional mean vector, Σ is the known covariance matrix and $\chi^2_k(p)$ is the quantile function for probability p of the chi-squared distribution with k degrees of freedom.[12] 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 (rate equal to half).

Complementary cumulative distribution function (tail distribution)

The complementary cumulative distribution function (ccdf) or the **tail distribution** is defined as $\overline{F}(\mathbf{x}) = 1 - \mathbb{P}(\mathbf{X} \leq \mathbf{x})$. When $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$, then the ccdf can be written as a probability the maximum of dependent Gaussian variables:[13]

$$\overline{F}(\mathbf{x}) = \mathbb{P}(\cup_i \{X_i \geq x_i\}) = \mathbb{P}(\max_i Y_i \geq 0), \quad \text{where } \mathbf{Y} \sim \mathcal{N}(\mu - \mathbf{x}, \Sigma).$$

While no simple closed formula exists for computing the ccdf, the maximum of dependent Gaussian variables can be estimated accurately via the Monte Carlo method.[13][14]

Properties

Probability in different domains

The probability content of the multivariate normal in a quadratic domain defined by $q(\boldsymbol{x}) = \boldsymbol{x}'\mathbf{Q_2}\boldsymbol{x} + \boldsymbol{q_1}'\boldsymbol{x} + q_0 > 0$ (where $\mathbf{Q_2}$ is a matrix, $\boldsymbol{q_1}$ is a vector, and q_0 is a scalar), which is relevant for Bayesian classification/decision theory using Gaussian discriminant analysis, is given by the generalized chi-squared distribution.[15] The probability content within any general domain defined by $f(\boldsymbol{x}) > 0$ (where $f(\boldsymbol{x})$ is a general function) can be computed using the numerical method of ray-tracing [15] (Matlab code (https://www.mathworks.com/matlabcentral/fileexchange/84973-integrate-and-classify-normal-distributions)).

Higher moments

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

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

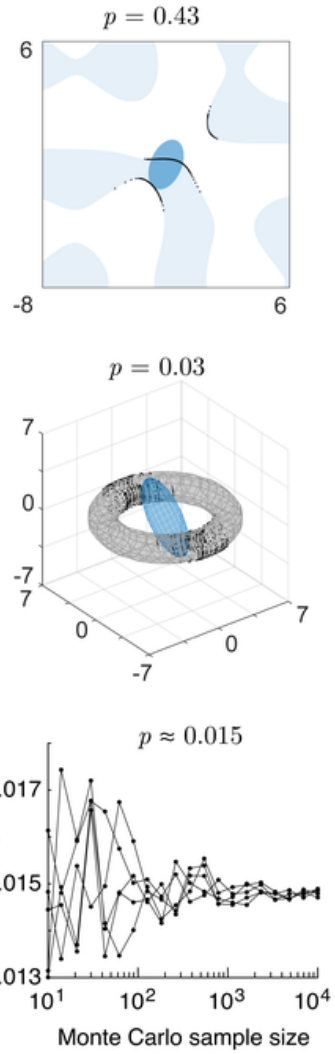
where $r_1 + r_2 + \cdots + r_N = k$.

The *k*th-order central moments are 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, for a *k*th $(= 2\lambda = 6)$ central moment, one sums the products of $\lambda = 3$ covariances (the expected value $\boldsymbol{\mu}$ is taken to be 0 in the interests of parsimony):



Top: the probability of a bivariate normal in the domain $\boldsymbol{x}\sin\boldsymbol{y} - \boldsymbol{y}\cos\boldsymbol{x} > 1$ (blue regions). Middle: the probability of a trivariate normal in a toroidal domain. Bottom: converging Monte-Carlo integral of the probability of a 4-variate normal in the 4d regular polyhedral domain defined by $\sum_{i=1}^{4} |\boldsymbol{x_i}| < 1$. These are all computed by the numerical method of ray-tracing. [15]

$$\begin{aligned}
&\mathrm{E}[X_1X_2X_3X_4X_5X_6] \\
&= \mathrm{E}[X_1X_2]\,\mathrm{E}[X_3X_4]\,\mathrm{E}[X_5X_6] + \mathrm{E}[X_1X_2]\,\mathrm{E}[X_3X_5]\,\mathrm{E}[X_4X_6] + \mathrm{E}[X_1X_2]\,\mathrm{E}[X_3X_6]\,\mathrm{E}[X_4X_5] \\
&\quad + \mathrm{E}[X_1X_3]\,\mathrm{E}[X_2X_4]\,\mathrm{E}[X_5X_6] + \mathrm{E}[X_1X_3]\,\mathrm{E}[X_2X_5]\,\mathrm{E}[X_4X_6] + \mathrm{E}[X_1X_3]\,\mathrm{E}[X_2X_6]\,\mathrm{E}[X_4X_5] \\
&\quad + \mathrm{E}[X_1X_4]\,\mathrm{E}[X_2X_3]\,\mathrm{E}[X_5X_6] + \mathrm{E}[X_1X_4]\,\mathrm{E}[X_2X_5]\,\mathrm{E}[X_3X_6] + \mathrm{E}[X_1X_4]\,\mathrm{E}[X_2X_6]\,\mathrm{E}[X_3X_5] \\
&\quad + \mathrm{E}[X_1X_5]\,\mathrm{E}[X_2X_3]\,\mathrm{E}[X_4X_6] + \mathrm{E}[X_1X_5]\,\mathrm{E}[X_2X_4]\,\mathrm{E}[X_3X_6] + \mathrm{E}[X_1X_5]\,\mathrm{E}[X_2X_6]\,\mathrm{E}[X_3X_4] \\
&\quad + \mathrm{E}[X_1X_6]\,\mathrm{E}[X_2X_3]\,\mathrm{E}[X_4X_5] + \mathrm{E}[X_1X_6]\,\mathrm{E}[X_2X_4]\,\mathrm{E}[X_3X_5] + \mathrm{E}[X_1X_6]\,\mathrm{E}[X_2X_5]\,\mathrm{E}[X_3X_4].
\end{aligned}$$

This yields $\frac{(2\lambda-1)!}{2^{\lambda-1}(\lambda-1)!}$ 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:

$$\begin{aligned}
\operatorname{E}\left[X_i^4\right] &= 3\sigma_{ii}^2 \\
\operatorname{E}\left[X_i^3 X_j\right] &= 3\sigma_{ii}\sigma_{ij} \\
\operatorname{E}\left[X_i^2 X_j^2\right] &= \sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2 \\
\operatorname{E}\left[X_i^2 X_j X_k\right] &= \sigma_{ii}\sigma_{jk} + 2\sigma_{ij}\sigma_{ik} \\
\operatorname{E}[X_i X_j X_k X_n] &= \sigma_{ij}\sigma_{kn} + \sigma_{ik}\sigma_{jn} + \sigma_{in}\sigma_{jk}.
\end{aligned}$$

where σ_{ij} is the covariance of X_i and X_j. With the above method one first finds the general case for a kth moment with k different X variables, $\boldsymbol{E}\left[X_i X_j X_k X_n\right]$, and then one simplifies this accordingly. For example, for $\operatorname{E}[X_i^2 X_k X_n]$, one lets $X_i = X_j$ and one uses the fact that $\sigma_{ii} = \sigma_i^2$.

Functions of a normal vector

A quadratic form of a normal vector $\boldsymbol{x}$, $q(\boldsymbol{x}) = \boldsymbol{x}'\mathbf{Q_2}\boldsymbol{x} + \boldsymbol{q_1}'\boldsymbol{x} + q_0$ (where $\mathbf{Q_2}$ is a matrix, $\boldsymbol{q_1}$ is a vector, and q_0 is a scalar), is a generalized chi-squared variable. [15]

If $\boldsymbol{f}(\boldsymbol{x})$ is a general scalar-valued function of a normal vector, its probability density function, cumulative distribution function, and inverse cumulative distribution function can be computed with the numerical method of ray-tracing (Matlab code (https://www.mathworks.com/matlabcentral/fileexchange/84973-integrate-and-classify-normal-distributions)).[15]

Likelihood function

If the mean and covariance matrix are known, the log likelihood of an observed vector $\boldsymbol{x}$ is simply the log of the probability density function:

$$\ln L(\boldsymbol{x}) = -\frac{1}{2}\left[\ln(|\boldsymbol{\Sigma}|) + (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + k\ln(2\pi)\right],$$

The circularly symmetric version of the noncentral complex case, where $\boldsymbol{z}$ is a vector of complex numbers, would be

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

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

A similar notation is used for multiple linear regression.[16]

Since the log likelihood of a normal vector is a quadratic form of the normal vector, it is distributed as a generalized chi-squared variable.[15]

Differential entropy

The differential entropy of the multivariate normal distribution is[17]

$$\begin{aligned}
h(f) &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\mathbf{x})\ln f(\mathbf{x})\,d\mathbf{x}, \\
&= \frac{1}{2}\ln(|(2\pi e)\,\boldsymbol{\Sigma}|) = \frac{1}{2}\ln\left((2\pi e)^k\,|\boldsymbol{\Sigma}|\right) = \frac{k}{2}\ln(2\pi e) + \frac{1}{2}\ln(|\boldsymbol{\Sigma}|) = \frac{k}{2} + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln(|\boldsymbol{\Sigma}|)
\end{aligned}$$

where the bars denote the matrix determinant and k is the dimensionality of the vector space.

Kullback–Leibler divergence

The Kullback–Leibler divergence from $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ to $\mathcal{N}_0(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, for non-singular matrices Σ_1 and Σ_0, is:[18]

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

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.

When $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$,

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

Mutual information

The mutual information of a distribution is a special case of the Kullback–Leibler divergence in which P is the full multivariate distribution and Q is the product of the 1-dimensional marginal distributions. In the notation of the Kullback–Leibler divergence section of this article, $\boldsymbol{\Sigma}_1$ is a diagonal matrix with the diagonal entries of $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$. The resulting formula for mutual information is:

$$I(\boldsymbol{X}) = -\frac{1}{2}\ln|\boldsymbol{\rho}_0|,$$

where $\boldsymbol{\rho}_0$ is the correlation matrix constructed from $\boldsymbol{\Sigma}_0$.

In the bivariate case the expression for the mutual information is:

$$I(x;y) = -\frac{1}{2}\ln(1-\rho^2).$$

Joint normality

Normally distributed and independent

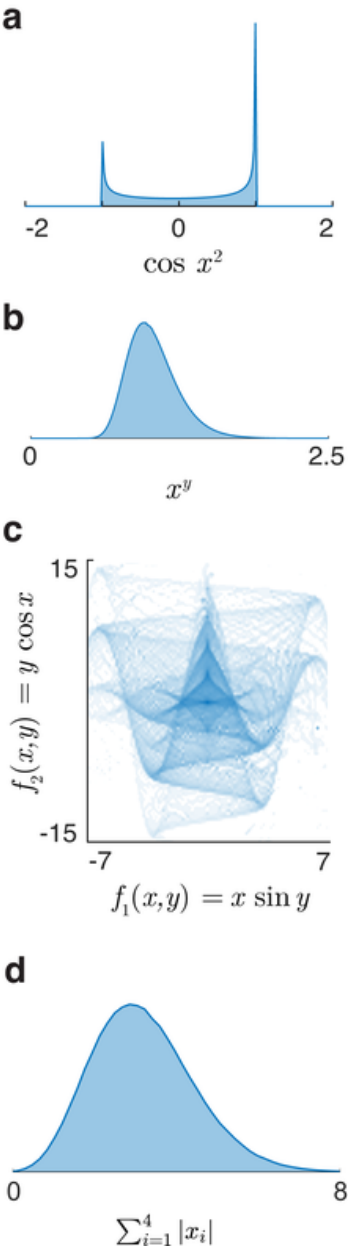
If X and Y are normally distributed and independent, this implies they are "jointly normally distributed", i.e., the pair (X, Y) must have multivariate normal distribution. However, a pair of jointly normally distributed variables need not be independent (would only be so if uncorrelated, $\rho = 0$).

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, as pointed out just above, it is *not* true that two random variables that are (*separately*, marginally) normally distributed and uncorrelated are independent.



a: Probability density of a function $\cos x^2$ of a single normal variable x with $\mu = -2$ and $\sigma = 3$. **b:** Probability density of a function x^y of a normal vector (x, y), with mean $\boldsymbol{\mu} = (1, 2)$, and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} .01 & .016 \\ .016 & .04 \end{bmatrix}$. **c:** Heat map of the joint probability density of two functions of a normal vector (x, y), with mean $\boldsymbol{\mu} = (-2, 5)$, and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} 10 & -7 \\ -7 & 10 \end{bmatrix}$. **d:** Probability density of a function $\sum_{i=1}^{4} |x_i|$ of 4 iid standard normal variables. These are computed by the numerical method of ray-tracing.[15]

Conditional distributions

If *N*-dimensional $\mathbf{x}$ 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 $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ 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 $\mathbf{x}_2 = \mathbf{a}$ is multivariate normal $(\mathbf{x}_1 \mid \mathbf{x}_2 = \mathbf{a}) \sim N(\bar{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$ where

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

and covariance matrix

$$\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$[19]

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 $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{a} - \boldsymbol{\mu}_2)$; compare this with the situation of not knowing the value of $\mathbf{a}$, in which case $\mathbf{x}_1$ would have distribution $\mathcal{N}_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

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[20]

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

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:

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

Proof: the result is obtained by taking the expectation of the conditional distribution $X_1 \mid X_2$ above.

In the centered case with 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

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

and the conditional variance is

$$\mathbf{var}(X_1 \mid X_2 = x_2) = 1 - \rho^2;$$

thus the conditional variance does not depend on x_2.

The conditional expectation of X_1 given that X_2 is smaller/bigger than z is:[21]:367

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

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

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

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

$\mathbf{E}(X_1 \mid X_2 < z) = \rho E(X_2 \mid X_2 < z)$ 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.[22]

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 $\mathbf{Y} = \mathbf{c} + \mathbf{B}\mathbf{X}$ is an affine transformation of $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{c}$ is an $M \times 1$ vector of constants and $\mathbf{B}$ is a constant $M \times N$ matrix, then $\mathbf{Y}$ has a multivariate normal distribution with expected value $\mathbf{c} + \mathbf{B}\boldsymbol{\mu}$ and variance $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\mathrm{T}}$ i.e., $\mathbf{Y} \sim \mathcal{N}\left(\mathbf{c} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^{\mathbf{T}}\right)$. 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 $\mathbf{Z} = \mathbf{b} \cdot \mathbf{X}$, where $\mathbf{b}$ is a constant vector with the same number of elements as $\mathbf{X}$ and the dot indicates the dot product, is univariate Gaussian with $Z \sim \mathcal{N}\left(\mathbf{b} \cdot \boldsymbol{\mu}, \mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}\mathbf{b}\right)$. This result follows by using

$$\mathbf{B} = [\, b_1 \quad b_2 \quad \dots \quad b_n \,] = \mathbf{b}^{\mathbf{T}}.$$

Observe how the positive-definiteness of $\boldsymbol{\Sigma}$ implies that the variance of the dot product must be positive.

An affine transformation of $\mathbf{X}$ such as $2\mathbf{X}$ is not the same as the sum of two independent realisations of $\mathbf{X}$.

Geometric interpretation

The equidensity contours of a non-singular multivariate normal distribution are ellipsoids (i.e. linear transformations of hyperspheres) centered at the mean.[23] 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 $\boldsymbol{\Sigma}$. The squared relative lengths of the principal axes are given by the corresponding eigenvalues.

If $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}(\mathbf{U}\boldsymbol{\Lambda}^{1/2})^{\mathrm{T}}$ is an eigendecomposition where the columns of $\mathbf{U}$ are unit eigenvectors and $\boldsymbol{\Lambda}$ 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, $\mathbf{U}$ 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 $\mathbf{U}$'s determinant. The distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is in effect $N(0, \mathbf{I})$ scaled by $\boldsymbol{\Lambda}^{1/2}$, rotated by $\mathbf{U}$ and translated by $\boldsymbol{\mu}$.

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; this is the degenerate case.

"The radius around the true mean in a bivariate normal random variable, re-written in polar coordinates (radius and angle), follows a Hoyt distribution."[24]

In one dimension the probability of finding a sample of the normal distribution in the interval $\mu \pm \sigma$ is approximately 68.27%, but in higher dimensions the probability of finding a sample in the region of the standard deviation ellipse is lower.[25]

Dimensionality	Probability
1	0.6827
2	0.3935
3	0.1987
4	0.0902
5	0.0374
6	0.0144
7	0.0052
8	0.0018
9	0.0006
10	0.0002

Statistical inference

Parameter estimation

The derivation of the maximum-likelihood estimator of the covariance matrix of a multivariate normal distribution is straightforward.

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

$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} = \frac{1}{n-1}\left[X'\left(I - \frac{1}{n}\cdot J\right)X\right]$ (matrix form; $\boldsymbol{I}$ is the $\boldsymbol{K}\times\boldsymbol{K}$ identity matrix, J is a $\boldsymbol{K}\times\boldsymbol{K}$ matrix of ones; the term in parentheses is thus the $\boldsymbol{K}\times\boldsymbol{K}$ centering matrix)

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

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

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

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

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[26] and Smith and Jain's adaptation[27] of the Friedman–Rafsky test created by Larry Rafsky and Jerome Friedman.[28]

Mardia's test[29] 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

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T \\ A &= \frac{1}{6n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\mathbf{x}_i - \bar{\mathbf{x}})^T\,\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})\right]^3 \\ B &= \sqrt{\frac{n}{8k(k+2)}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{x}_i - \bar{\mathbf{x}})^T\,\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\right]^2 - k(k+2)\right\} \end{aligned}$$

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 $(50 \le n < 400)$, the parameters of the asymptotic distribution of the kurtosis statistic are modified[30] For small sample tests ($n < 50$) empirical critical values are used. Tables of critical values for both statistics are given by Rencher[31] 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.[32]

The **BHEP test**[33] 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

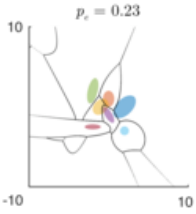
$$T_\beta = \int_{\mathbb{R}^k} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\mathbf{t}^T \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}})} - e^{-|\mathbf{t}|^2/2} \right|^2 \mu_\beta(\mathbf{t})\, d\mathbf{t}$$
$$= \frac{1}{n^2} \sum_{i,j=1}^{n} e^{-\frac{\beta^2}{2}(\mathbf{x}_i - \mathbf{x}_j)^T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

The limiting distribution of this test statistic is a weighted sum of chi-squared random variables,[33] 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.[34]

Classification into multivariate normal classes

Gaussian Discriminant Analysis



	1	2	3	4	5	6	7
1	0.1024	0.0114	0.0019	0.0181	4.8412e-05	0.0058	0.0032
2	0.0074	0.1021	0.0119	0.0157	0.0056	1.4404e-05	1.2221e-04
3	0.0028	0.0340	0.0626	0.0225	0.0160	6.9137e-06	0.0051
4	0.0029	0.0094	0.0180	0.1015	0.0018	0.0074	0.0019
5	7.6722e-06	0.0040	0.0142	6.7751e-05	0.1233	1.1630e-12	0.0013
6	0.0017	1.2153e-09	8.3067e-10	0.0032	8.0951e-21	0.1379	3.7230e-05
7	1.3349e-04	1.1911e-14	0.0011	0.0030	5.7880e-05	1.9171e-04	0.1384

Left: Classification of seven multivariate normal classes. Coloured ellipses are 1 sd error ellipses. Black marks the boundaries between the classification regions. p_e is the probability of total classification error. Right: the error matrix. p_{ij} is the probability of classifying a sample from normal i as j. These are computed by the numerical method of ray-tracing [15] (Matlab code (https://www.mathworks.com/matlabcentral/fileexchange/84973-integrate-and-classify-normal-distributions)).

Suppose that observations (which are vectors) are presumed to come from one of several multivariate normal distributions, with known means and covariances. Then any given observation can be assigned to the distribution from which it has the highest probability of arising. This classification procedure is called Gaussian discriminant analysis. The classification performance, i.e. probabilities of the different classification outcomes, and the overall classification error, can be computed by the numerical method of ray-tracing [15] (Matlab code (https://www.mathworks.com/matlabcentral/fileexchange/84973-integrate-and-classify-normal-distributions)).

Computational methods

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:[35]

1. Find any real matrix **A** such that **A** $\mathbf{A}^T = \boldsymbol{\Sigma}$. 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 $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ 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 $\mathbf{z} = (z_1, \ldots, z_N)^T$ 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 (Euclidean norm or vector length) of a multivariate normally distributed vector (uncorrelated and zero centered).
  - Rayleigh distribution, the pdf of the vector length of a bivariate normally distributed vector (uncorrelated and zero centered)
  - Rice distribution, the pdf of the vector length of a bivariate normally distributed vector (uncorrelated and non-centered)
  - Hoyt distribution, the pdf of the vector length of a bivariate normally distributed vector (correlated and centered)
- Complex normal distribution, an application of bivariate normal distribution
- Copula, for the definition of the Gaussian or normal copula model.
- Multivariate t-distribution, which is another widely used spherically symmetric multivariate distribution.
- Multivariate stable distribution extension of the multivariate normal distribution, when the index (exponent in the characteristic function) is between zero and two.
- Mahalanobis distance
- Wishart distribution
- Matrix normal distribution

References

1. Lapidoth, Amos (2009). *A Foundation in Digital Communication*. Cambridge University Press. ISBN 978-0-521-19395-5.
2. Gut, Allan (2009). *An Intermediate Course in Probability*. Springer. ISBN 978-1-441-90161-3.
3. Kac, M. (1939). "On a characterization of the normal distribution". *American Journal of Mathematics*. **61** (3): 726–728. doi:10.2307/2371328 (https://doi.org/10.2307%2F2371328). JSTOR 2371328 (https://www.jstor.org/stable/2371328).
4. Sinz, Fabian; Gerwinn, Sebastian; Bethge, Matthias (2009). "Characterization of the p-generalized normal distribution" (https://doi.org/10.1016%2Fj.jmva.2008.07.006). *Journal of Multivariate Analysis*. **100** (5): 817–820. doi:10.1016/j.jmva.2008.07.006 (https://doi.org/10.1016%2Fj.jmva.2008.07.006).
5. Simon J.D. Prince(June 2012). Computer Vision: Models, Learning, and Inference (http://www.computervisionmodels.com/). Cambridge University Press. 3.7:"Multivariate normal distribution".
6. 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://doi.org/10.2307%2F2318494). JSTOR 2318494 (https://www.jstor.org/stable/2318494).
7. Wyatt, John (November 26, 2008). "Linear least mean-squared error estimation" (https://web.archive.org/web/20151010114443/http://web.mit.edu/6.041/www/LECTURE/lec22.pdf) (PDF). *Lecture notes course on applied probability*. Archived from the original (http://web.mit.edu/6.041/www/LECTURE/lec22.pdf) (PDF) on October 10, 2015. Retrieved 23 January 2012.
8. Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley. pp. 527–528. ISBN 0-471-70823-2.
9. Botev, Z. I. (2016). "The normal law under linear restrictions: simulation and estimation via minimax tilting". *Journal of the Royal Statistical Society, Series B*. **79**: 125–148. arXiv:1603.04166 (https://arxiv.org/abs/1603.04166). Bibcode:2016arXiv160304166B (https://ui.adsabs.harvard.edu/abs/2016arXiv160304166B). doi:10.1111/rssb.12162 (https://doi.org/10.1111%2Frssb.12162). S2CID 88515228 (https://api.semanticscholar.org/CorpusID:88515228).
10. Genz, Alan (2009). *Computation of Multivariate Normal and t Probabilities* (https://www.springer.com/statistics/computational+statistics/book/978-3-642-01688-2). Springer. ISBN 978-3-642-01689-9.
11. 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)
12. 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://doi.org/10.1007%2FBF02868568). S2CID 123269490 (https://api.semanticscholar.org/CorpusID:123269490).
13. Botev, Z. I.; Mandjes, M.; Ridder, A. (6–9 December 2015). "Tail distribution of the maximum of correlated Gaussian random variables". *2015 Winter Simulation Conference (WSC)*. Huntington Beach, Calif., USA: IEEE. pp. 633–642. doi:10.1109/WSC.2015.7408202 (https://doi.org/10.1109%2FWSC.2015.7408202). ISBN 978-1-4673-9743-8.
14. Adler, R. J.; Blanchet, J.; Liu, J. (7–10 Dec 2008). "Efficient simulation for tail probabilities of Gaussian random fields". *2008 Winter Simulation Conference (WSC)*. Miami, Fla., USA: IEEE. pp. 328–336. doi:10.1109/WSC.2008.4736085 (https://doi.org/10.1109%2FWSC.2008.4736085). ISBN 978-1-4244-2707-9.
15. Das, Abhranil (2020). "A method to integrate and classify normal distributions". arXiv:2012.14331 (https://arxiv.org/abs/2012.14331) [stat.ML (https://arxiv.org/archive/stat.ML)].
16. Tong, T. (2010) Multiple Linear Regression : MLE and Its Distributional Results (http://amath.colorado.edu/courses/7400/2010Spr/lecture9.pdf) Archived (https://www.webcitation.org/6HPbX5thy?url=http://amath.colorado.edu/courses/7400/2010Spr/lecture9.pdf) 2013-06-16 at WebCite, Lecture Notes
17. Gokhale, DV; Ahmed, NA; Res, BC; Piscataway, NJ (May 1989). "Entropy Expressions and Their Estimators for Multivariate Distributions". *IEEE Transactions on Information Theory*. **35** (3): 688–692. doi:10.1109/18.30996 (https://doi.org/10.1109%2F18.30996).

18. Duchi, J. "Derivations for Linear Algebra and Optimization" (https://stanford.edu/~jduchi/projects/general_notes.pdf#page=13) (PDF): 13.
19. Eaton, Morris L. (1983). *Multivariate Statistics: a Vector Space Approach*. John Wiley and Sons. pp. 116–117. ISBN 978-0-471-02776-8.
20. Jensen, J (2000). *Statistics for Petroleum Engineers and Geoscientists*. Amsterdam: Elsevier. p. 207.
21. Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press. ISBN 0-521-33825-5.
22. An algebraic computation of the marginal distribution is shown here http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html Archived (https://web.archive.org/web/20100117200722/http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html) 2010-01-17 at the Wayback Machine. A much shorter proof is outlined here https://math.stackexchange.com/a/3832137
23. Nikolaus Hansen (2016). "The CMA Evolution Strategy: A Tutorial" (https://web.archive.org/web/20100331114258/http://www.lri.fr/~hansen/cmatutorial.pdf) (PDF). arXiv:1604.00772 (https://arxiv.org/abs/1604.00772). Bibcode:2016arXiv160400772H (https://ui.adsabs.harvard.edu/abs/2016arXiv160400772H). Archived from the original (http://www.lri.fr/~hansen/cmatutorial.pdf) (PDF) on 2010-03-31. Retrieved 2012-01-07.
24. Daniel Wollschlaeger. "The Hoyt Distribution (Documentation for R package 'shotGroups' version 0.6.2)" (http://finzi.psych.upenn.edu/usr/share/doc/library/shotGroups/html/hoyt.html).
25. Wang, Bin; Shi, Wenzhong; Miao, Zelang (2015-03-13). Rocchini, Duccio (ed.). "Confidence Analysis of Standard Deviational Ellipse and Its Extension into Higher Dimensional Euclidean Space" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4358977). *PLOS ONE*. **10** (3): e0118537. Bibcode:2015PLoSO..1018537W (https://ui.adsabs.harvard.edu/abs/2015PLoSO..1018537W). doi:10.1371/journal.pone.0118537 (https://doi.org/10.1371%2Fjournal.pone.0118537). ISSN 1932-6203 (https://www.worldcat.org/issn/1932-6203). PMC 4358977 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4358977). PMID 25769048 (https://pubmed.ncbi.nlm.nih.gov/25769048).
26. Cox, D. R.; Small, N. J. H. (1978). "Testing multivariate normality". *Biometrika*. **65** (2): 263. doi:10.1093/biomet/65.2.263 (https://doi.org/10.1093%2Fbiomet%2F65.2.263).
27. 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://doi.org/10.1109%2F34.6789).
28. Friedman, J. H.; Rafsky, L. C. (1979). "Multivariate Generalizations of the Wald–Wolfowitz and Smirnov Two-Sample Tests" (https://doi.org/10.1214%2Faos%2F1176344722). *The Annals of Statistics*. **7** (4): 697. doi:10.1214/aos/1176344722 (https://doi.org/10.1214%2Faos%2F1176344722).
29. 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://doi.org/10.1093%2Fbiomet%2F57.3.519).
30. Rencher (1995), pages 112–113.
31. Rencher (1995), pages 493–495.
32. Baringhaus, L.; Henze, N. (1991). "Limit distributions for measures of multivariate skewness and kurtosis based on projections" (https://doi.org/10.1016%2F0047-259X%2891%2990031-V). *Journal of Multivariate Analysis*. **38**: 51–69. doi:10.1016/0047-259X(91)90031-V (https://doi.org/10.1016%2F0047-259X%2891%2990031-V).
33. 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://doi.org/10.1007%2FBF02613322). S2CID 122362448 (https://api.semanticscholar.org/CorpusID:122362448).
34. 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://doi.org/10.1007%2Fs00362-002-0119-6). S2CID 122934510 (https://api.semanticscholar.org/CorpusID:122934510).
35. Gentle, J. E. (2009). *Computational Statistics* (http://cds.cern.ch/record/1639470). Statistics and Computing. New York: Springer. pp. 315–316. doi:10.1007/978-0-387-98144-4 (https://doi.org/10.1007%2F978-0-387-98144-4). ISBN 978-0-387-98143-7.

Literature

- Rencher, A.C. (1995). *Methods of Multivariate Analysis*. New York: Wiley.
- Tong, Y. L. (1990). *The multivariate normal distribution*. Springer Series in Statistics. New York: Springer-Verlag. doi:10.1007/978-1-4613-9655-0 (https://doi.org/10.1007%2F978-1-4613-9655-0). ISBN 978-1-4613-9657-4.