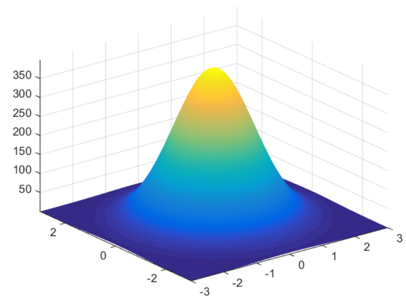
Multivariate Normal Distribution

A **multivariate normal distribution** is a vector in multiple normally distributed variables, such that any linear combinatio variables is also normally distributed. It is mostly useful in extending the central limit theorem to multiple variables, but applications to bayesian inference and thus machine learning, where the multivariate normal distribution is used to approximate the features of some characteristics; for instance, in detecting faces in pictures.



A normal (or Gaussian) distribution in 2 variables

Contents

Formal definitions

A **random vector** $\mathbf{x} = (X_1, X_2, \ldots, X_n)$ is multivariate normal if any linear combination of the random variables $X_1, X_2, \ldots, X_n$ is normally distributed. In other words,

$$a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

has a normal distribution for any constants $a_1, a_2, \ldots, a_n$.

Equivalently, multivariate distributions can be viewed as a linear transformation of a collection of independent standard random variables, meaning that if $\mathbf{z}$ is another random vector whose components are all standard random variables, the a matrix A and vector μ such that

$$\mathbf{x} = A\mathbf{z} + \mu.$$

If $\mathbf{x}$ is multivariate normal, it also has a **mean vector** μ such that

$$\mu = (\mathbb{E}(X_1), \mathbb{E}(X_2), \ldots, \mathbb{E}(X_n))$$

where $\mathbb{E}(X)$ is the expected value (or mean) of X. $\mathbf{x}$ also has a **covariance matrix** Σ satisfying

$$\Sigma_{i,j} = \operatorname{Cov}(X_i, X_j)$$

where $\operatorname{Cov}(X_i, X_j)$ is the covariance of X_i and X_j.

Furthermore, $\mathbf{x}$ is completely defined by μ and Σ, so it is convenient to write

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

Properties

The p.d.f. of a multivariate normal distribution $\mathbf{x}$ is given by

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right)$$

where $\exp(x) = e^x$.

In understanding this equation, it is useful to compare with the p.d.f. of the univariate normal distribution, or the normal distribution in 1 variable, which is given by

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

In this univariate case, $-\frac{1}{2\sigma^2}(x-\mu)^2$ is a quadratic function of x, which is a parabola that opens downward due to the r leading coefficient. In the multivariate case, $-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)$ is a quadratic form in the vector $\mathbf{x}$. Since Σ is pos definite, this quadratic form is negative definite, and so opens a "bowl" oriented downward in an analogous way to how parabola in the univariate case opens downwards.

The leading coefficient in the univariate case $\frac{1}{\sqrt{2\pi}\sigma}$ does not depend on x, and is chosen in such a way that

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = 1$$

Similarly, the leading coefficient in the multivariate case $\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}}$ does not depend on $\mathbf{x}$, and is chosen in such a way th

$$\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right) = 1$$

It is also worth noting that the multivariate formula reduces to the univariate one in the case $n = 1$, as in this case $(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x}-\mu) = (x-\mu)\frac{1}{\sigma^2}(x-\mu)$.

One main importance of the multivariate distribution is an extension of the central limit theorem to multiple variables:

THEOREM

Suppose $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of independent, identically distributed random vectors with common mean vector μ positive-definite \covariance matrix σ. Then

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu)$$

converges to

$$Y_n \sim \mathcal{N}(0, \Sigma)$$

Applications

The multivariate normal distribution is useful in analyzing the relationship between multiple normally distributed variabl thus has heavy application to biology and economics where the relationship between approximately-normal variables is interest. For instance, one of the earliest uses of the multivariate distribution was in analyzing the relationship between a father's height and the height of their eldest son, resolving a question Darwin posed in *On the Origin of Species*. That w revealed that:

- Both fathers' and son' height were normally distributed with mean 68 and variance 3 (in inches).
- For any given height f, the height of sons whose fathers were of height f was *also* normally distributed, and in fact th average height was a linear function of f.

In essence, Galton had discovered the **conditional distribution** of the multivariate normal: if $\mathbf{x}$ is partitioned into $\begin{pmatrix}\mathbf{x_1}\\ \mathbf{x_2}\end{pmatrix}$, $\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$, and Σ into $\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma'_{12} & \Sigma_{22}\end{pmatrix}$, then

$$\mathbf{X_1}|\mathbf{X_2} \sim \mathcal{N}(\mu_1 + \sigma_{12}\sigma_{22}^{-1}(\mathbf{X_2} - \mu_2), \sigma_{11} - \sigma_{12}\sigma_{22}^{-1}\sigma'_{12})$$

whose importance mainly extends to the fact that the conditional distribution $\mathbf{x}|\mathbf{y}$ of two multivariate normal variables is normal distribution, where the parameters are functions of $\mathbf{y}$.

In modern time, the multivariate normal distribution is incredibly important in **machine learning**, whose purpose is (very roughly speaking) to categorize input data x into labels y, based on some training pairs x, y. One major approach invol analyzing the distribution $p(x|y)$, and approximating it with a multivariate normal distribution, the validity of which can checked using various **normality tests**; paradoxically, however, classifying based on multivariate normal distributions has successful in practice even when it is known to be a poor model for the data.

See Also

- Continuous random variables - probability density function (pdf)
- Normal distribution
- Covariance

Cite as: Multivariate Normal Distribution. *Brilliant.org*. Retrieved 09:24, January 19, 2020, from https://brilliant.org/wiki/multivariate-normal-distribution/