Note 18. Jointly Gaussian random variables

Alex Fu

Fall 2022

1 Definitions

One of the most common multivariate or vector-valued distributions is the natural generalization of the normal distribution $\mathcal{N}(\mu, \sigma^2)$, which shares one of its central properties.

Theorem 1 (Multivariate central limit theorem*).

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be i.i.d. random vectors with mean $\boldsymbol{\mu}$ and covariance matrix Σ, and let the sample mean be $\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$. Then

$$\sqrt{n}\left(\bar{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\text{d}} \mathcal{N}(0, \Sigma)$$

describes convergence in distribution to the multivariate normal distribution.

Definition 1 (Multivariate normal distribution, or jointly Gaussian random variables I).

Let $\mathbf{Z} \in \mathbb{R}^m$ be the *standard normal random vector*, whose entries Z_i are i.i.d. $\mathcal{N}(0, 1)$. Then $\mathbf{X} \in \mathbb{R}^n$ follows the **multivariate normal distribution** if it is an *affine combination* of $\mathbf{Z}$: there exist $A \in \mathbb{R}^{n \times m}$ and $\boldsymbol{\mu} \in \mathbb{R}^n$ such that

$$\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}.$$

If $\mathbf{X}$ follows the multivariate normal distribution, we say that the entries of $\mathbf{X}$, the random variables $X_1, \ldots, X_n$, are **jointly Gaussian**.

The mean of $\mathbf{X}$ is the random vector $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$, the covariance (matrix) of $\mathbf{X}$ is $\text{var}(\mathbf{X}) = \Sigma = AA^{\mathsf{T}}$, and the distribution of $\mathbf{X}$ is denoted $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The multivariate normal distribution is uniquely parametrized by its mean and covariance.

Let us verify the mean and covariance of $\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}$. By the linearity of expectation,

$$\mathbb{E}(\mathbf{X}) = A\,\mathbb{E}(\mathbf{Z}) + \boldsymbol{\mu} = \boldsymbol{\mu}$$

$$\text{var}(\mathbf{X}) = \mathbb{E}\big((\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\mathsf{T}\big) = \mathbb{E}\big((A\mathbf{Z})(A\mathbf{Z})^\mathsf{T}\big) = A\,\mathbb{E}\big(\mathbf{Z}\mathbf{Z}^\mathsf{T}\big)A^\mathsf{T} = AA^\mathsf{T}.$$

Conversely, given any covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, which must be *positive semidefinite*, we can find its square root $A \in \mathbb{R}^{n\times m}$ such that $AA^\mathsf{T} = \Sigma$.

By the spectral theorem, the symmetric matrix Σ can be decomposed as $U\Lambda U^\mathsf{T}$, where U is an orthonormal matrix of its eigenvectors and Λ is a diagonal matrix of its nonnegative eigenvalues. Λ also admits a square root $\Lambda^{1/2}$, given by $(\Lambda^{1/2})_{i,j} = (\Lambda_{i,j})^{1/2}$. Then a possible square root of Σ is $A = U\Lambda^{1/2}U^\mathsf{T}$, or $A = U\Lambda^{1/2}$.

The square root $A = U\Lambda^{1/2}$ also gives rise to a geometric interpretation:

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

so every $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is simply the multivariate standard normal distribution $\mathcal{N}(\mathbf{0}, I)$ scaled by $\Lambda^{1/2}$, rotated by U, and translated by $\boldsymbol{\mu}$. In other words, every $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is an affine transformation of $\mathcal{N}(\mathbf{0}, I)$, just as every $\mathcal{N}(\mu, \sigma^2)$ is an affine transformation of $\mathcal{N}(0, 1)$.

Definition 2 (Multivariate normal distribution, or jointly Gaussian random variables II).

Equivalently, the random variables $X_1, \ldots, X_n$ are **jointly Gaussian** if any linear combination of them follows the (univariate) Gaussian distribution:

$$\mathbf{c}^\mathsf{T}\mathbf{X} = c_1X_1 + \cdots + c_nX_n \sim \mathcal{N}(\mu, \sigma^2) \text{ for some } \mu, \sigma^2 \in \mathbb{R}.$$

Let us show that Definition 2 is equivalent to Definition 1. Given $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$,

$$\mathbf{c}^\mathsf{T}\mathbf{X} = \mathbf{c}^\mathsf{T}(A\mathbf{Z} + \boldsymbol{\mu}) = \sum_{i=1}^{n} c_i(\text{row}_i(A)\mathbf{Z} + \mu_i) = \sum_{i=1}^{n} c_i\mu_i + \sum_{i=1}^{n}\sum_{j=1}^{m} c_iA_{i,j}Z_j$$

is an affine combination of the independent $Z_i \sim \mathcal{N}(0, 1)$, which we know is normally distributed as well. Conversely, given jointly Gaussian $X_1, \ldots, X_n$, we will need the following result.

Proposition 1 (Moment-generating function of the multivariate normal distribution*).

The joint moment-generating function of $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is

$$M_\mathbf{X}(\mathbf{t}) = \mathbb{E}\big(\exp(\mathbf{t}^\mathsf{T}\mathbf{X})\big) = \exp\left(\mathbf{t}^\mathsf{T}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^\mathsf{T}\Sigma\mathbf{t}\right).$$

Proof that Definition 2 implies Definition 1. Let $Y = \mathbf{c}^\mathsf{T}\mathbf{X}$ be distributed as $\mathcal{N}(\mu, \sigma^2)$, where

$$\begin{aligned}\mu &= \mathbf{c}^\mathsf{T}\,\mathbb{E}(\mathbf{X})\\ \sigma^2 &= \mathbf{c}^\mathsf{T}\,\mathbb{E}\big((\mathbf{X}-\mathbb{E}(\mathbf{X}))(\mathbf{X}-\mathbb{E}(\mathbf{X}))^\mathsf{T}\big)\mathbf{c}.\end{aligned}$$

Then, we observe that $M_{\mathbf{X}}(\mathbf{c}) = \mathbb{E}\big(\exp(\mathbf{c}^\mathsf{T}\mathbf{X})\big)$ is precisely equal to

$$\begin{aligned} M_Y(1) = \mathbb{E}(\exp(Y)) &= \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ &= \exp\left(\mathbf{c}^\mathsf{T}\,\mathbb{E}(\mathbf{X}) + \frac{1}{2}\mathbf{c}^\mathsf{T}\,\mathbb{E}\big((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\mathsf{T}\big)\mathbf{c}\right), \end{aligned}$$

the moment-generating function of a multivariate normal distribution. □

Proposition 2 (Dot product of multivariate normal distribution).

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then the dot product $\mathbf{u}^\mathsf{T}\mathbf{X}$ is distributed as $\mathcal{N}(\mathbf{u}\boldsymbol{\mu}, \mathbf{u}^\mathsf{T}\Sigma\mathbf{u})$. In particular, if $\mathbf{u}$ is a unit vector, then $(\mathbf{u}^\mathsf{T}\mathbf{X})\mathbf{u}$ is the projection of $\mathbf{X}$ onto the direction of $\mathbf{u}$.

2 Properties

Affine combinations are a key part of both definitions of jointly Gaussian random variables, so they unsurprisingly turn out to be involved in key properties as well.

Proposition 3 (Family of distributions closed under affine combinations).

Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, let $A \in \mathbb{R}^{\ell \times n}$, and let $\mathbf{b} \in \mathbb{R}^\ell$. Then the affine combination $A\mathbf{X} + \mathbf{b}$ is also normally distributed as $\mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^\mathsf{T})$.

Proof. By definition, $\mathbf{X} = \Sigma^{1/2}\mathbf{Z} + \boldsymbol{\mu}$ for some standard normal random vector $\mathbf{Z} \in \mathbb{R}^m$, so

$$\mathbf{Y} = A\mathbf{X} + \mathbf{b} = (A\Sigma^{1/2})\mathbf{Z} + (A\boldsymbol{\mu} + \mathbf{b})$$

is also normally distributed. □

In particular, linear combinations of jointly Gaussian random variables are jointly Gaussian as well. Another key property of the multivariate normal distribution is that the first and second moments are sufficient information to fully determine it. The moment-generating function was uniquely determined by $\boldsymbol{\mu}$ and Σ, and so is its inverse Laplace transform, the probability density function.

Proposition 4 (Probability density function of the multivariate normal distribution*).

Recall that Σ is positive semidefinite, so Σ is invertible iff it is positive definite. Then the pdf of $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, which exists iff Σ is positive definite, is equal to

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

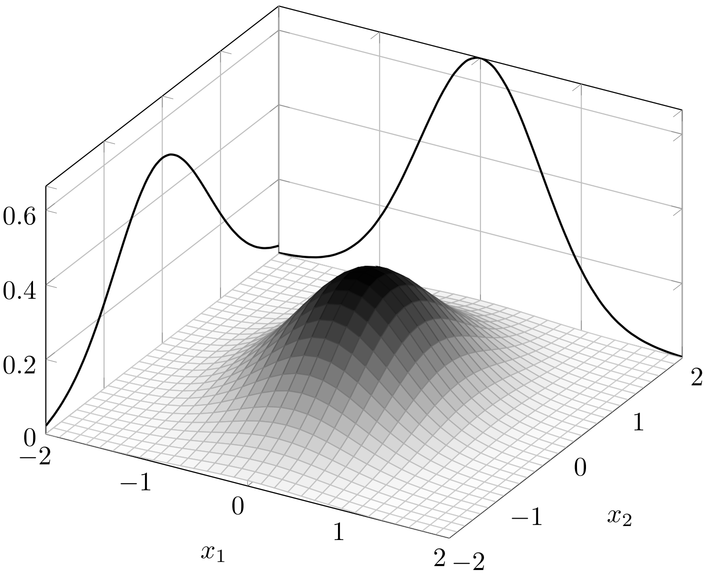

Figure 1: The density of a 2-dimensional multivariate normal distribution.

We note that the density function above forms a figure encompassing 3-dimensional volume. If Σ is noninvertible, so that its determinant is zero, and one or more dimensions are collapsed, then the figure formed is *degenerate*: even if it may have nonzero area, it has zero volume. Then there is no meaningful density for "probability mass 1 per zero volume."

We can also see that the projections onto the individual axes, the *marginal* distributions X_i, are themselves 1-dimensional Gaussian distributions. In fact, any vertical "slice" of the figure in any direction, such as along the line $x_1 + 2x_2 = 5$, will be 1-dimensional Gaussian, which describes that any affine combination of the X_i is normal — Definition 2.

> "Finding an affine combination of X_i is like taking a picture of a mountain from the side."

The intersection of $f_{\mathbf{X}}$ with any vertical plane is Gaussian, so what about its intersection with a horizontal plane? We observe that the significant term in the pdf is

$$g(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

The **level curves** of g are the sets of points which have the same value of g, which means that they have equal probability density: the level curves of g are the level curves of $f_{\mathbf{X}}$.

Proposition 5 (Level curves of $f_{\mathbf{X}}$*).

The level curves of the pdf are *hyperellipsoids*, multidimensional generalizations of ellipses.

1. In the case of $\Sigma = I$, the level curves of

   $$g(\mathbf{x}) = \mathbf{x}^{\mathsf{T}} \Sigma^{-1} \mathbf{x} = \|\mathbf{x}\|_2^2$$

   are circles, or more generally hyperspheres, which have constant radius, centered at $\boldsymbol{\mu}$.

2. In the case of $\Sigma = \Lambda$, a positive diagonal matrix, the level curves of

$$g(\mathbf{x}) = \mathbf{x}^\mathsf{T}\Lambda^{-1}\mathbf{x} = \sum_{i=1}^{n} \frac{1}{\lambda_i} x_i^2$$

are now hyperellipsoids centered at $\boldsymbol{\mu}$, whose axes are parallel to the coordinate axes, and whose semiaxis length in the ith coordinate direction is $(\lambda_i)^{1/2}$.

3. In the general case, where $\Sigma = U\Lambda U^\mathsf{T}$, the level curves of

$$g(\mathbf{x}) = (U^\mathsf{T}\mathbf{x})^\mathsf{T}\Lambda^{-1}U^\mathsf{T}\mathbf{x} = \sum_{i=1}^{n} \frac{1}{\lambda_i}(U^\mathsf{T}\mathbf{x})_i^2$$

are again hyperellipsoids with the same semiaxis lengths, but whose axes are in the directions given by the columns of U.

We again note that the semiaxis lengths $(\lambda_i)^{1/2}$ are given by scaling, the axis directions $\mathrm{col}_i(U)$ given by rotation, the center $\boldsymbol{\mu}$ given by translation, and the "standard" level curves given by hyperspheres centered at the origin, analogously to how $\mathcal{N}(\boldsymbol{\mu}, \Sigma) = \boldsymbol{\mu} + U\Lambda^{1/2}\mathcal{N}(0, I)$.

The following point deserves emphasis.

Proposition 6.

Jointly Gaussian implies marginally Gaussian, but not the converse.

Only when the *joint* distribution of $X_1, \ldots, X_n$ is Gaussian do these properties apply, not when $X_1, \ldots, X_n$ is any collection of marginally Gaussian random variables, which can be related in arbitrary ways that most often cannot be summarized by only the first two moments.

Proposition 7 (Independent iff uncorrelated).

Jointly Gaussian random variables are independent iff they are uncorrelated.

Proof. Independence implies uncorrelatedness in general, so we show the converse, though only for positive definite Σ. As $\mathrm{cov}(X_i, X_j) = 0$ for all $i \neq j$, the covariance Σ is diagonal. Then

$$\begin{aligned}
f_\mathbf{X}(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^n}\left(\prod_{i=1}^n \sigma_i\right)} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu_i)^2}{\sigma_i^2}\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2}\frac{(x_i - \mu_i)^2}{\sigma_i^2}\right) \\
&= \prod_{i=1}^{n} f_{X_i}(x_i).
\end{aligned}$$

□

Example 1 (Uncorrelated but not independent marginally Gaussian random variables).

Let $X \sim \mathcal{N}(0,1)$, and let $Y = WX$, where W is Rademacher and independent of X. Then X and Y are uncorrelated,

$$\operatorname{cov}(X,Y) = \mathbb{E}(WX^2) - \mathbb{E}(X)\,\mathbb{E}(WX) = \mathbb{E}(W)\,\mathbb{E}(X^2) - 0 = 0,$$

but not independent, as

$$\mathbb{P}(X \le -1 \mid Y = 0) = 0 \ne \mathbb{P}(X \le -1).$$

As a consequence of Proposition 7, we find another simplification in estimation, where the LLSE only involves entries in $\boldsymbol{\mu}$ and Σ.

Proposition 8 (MMSE and LLSE are equivalent).

For jointly Gaussian random variables, $\mathbb{E}(X \mid Y) = \mathbb{L}(X \mid Y)$.

Proof. In general, the MMSE is only equal to the LLSE if it is affine, so we will need to use the special property that uncorrelatedness implies independendence.

By the orthogonality principle, $X - \mathbb{L}(X \mid Y)$ is zero-mean and orthogonal to Y, which implies that $\operatorname{cov}(X - \mathbb{L}(X \mid Y), Y) = 0$. As both are affine combinations of the jointly Gaussian X and Y, they are thus also independent.

Functions of independent random variables are independent, so $X - \mathbb{L}(X \mid Y)$ is independent of every $\varphi(Y)$, which in turn implies orthogonality. Therefore $\mathbb{E}(X \mid Y) = \mathbb{L}(X \mid Y)$. □

In summary, the multivariate normal distribution for random vectors, whose entries are *jointly Gaussian*, enjoys a special affinity with affine combinations and a unique characterization by the first two moments, which reduces properties of the distribution such as level curves of the PDF, independence, and the MMSE to simpler statements about the mean and covariance.

■