Jointly Gaussian

EECS 126 (UC Berkeley)

Spring 2022

1 Introduction

1.1 Definitions

We list equivalent definitions of jointly Gaussian random variables below.

Definition 1. Let random vector $\mathbf{X} := (X_1, \dots, X_n)^\top$. Let $\mathbf{Z} \in \mathbb{R}^\ell$ be the standard normal random vector (i.e. $Z_i \sim \mathcal{N}(0,1)$ for $i = 1, \dots, \ell$ are i.i.d.). Then $X_1, \dots, X_n$ are *jointly Gaussian* if there exist $\boldsymbol{\mu} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times \ell}$ such that $\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}$.

Definition 2. $X_1, \dots, X_n$ are *jointly Gaussian* if any linear combination of $X_1, \dots, X_n$, $u^\top \mathbf{X}$, follows a normal distribution.

1.2 Probability Density Function

Given a positive definite Σ, the joint PDF of $\mathbf{X}$ is

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

1.3 Proof of Covariance Matrix Expression

Finally, let's prove that $\Sigma = AA^\top$.

Proof. For any general random vector $\mathbf{X}$, we define the (i,j)-entry of the square covariance matrix $\Sigma_{\mathbf{XX}}$ as $\Sigma_{i,j} = \operatorname{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$. Let $\mathbf{X}$ be jointly Gaussian. By definition 1, $\mathbf{X} = A\mathbf{Z} + \boldsymbol{\mu}$. We thus see that

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[A\mathbf{Z} + \boldsymbol{\mu}] = A\,\mathbb{E}[\mathbf{Z}] + \mathbb{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}.$$

Hence,

$$\Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top]$$

$$
\begin{aligned}
&= \mathbb{E}[(A\mathbf{Z})(A\mathbf{Z})^\top] \\
&= \mathbb{E}[A\mathbf{Z}\mathbf{Z}^\top A^\top] \\
&= A\,\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]A^\top \\
&= AA^\top
\end{aligned}
$$

□

2 Properties of JG RVs

2.1 Independent iff Uncorrelated

In general, for any two random variables X_1, X_2, if X_1 and X_2 are independent, then they are necessarily uncorrelated:

$$\text{cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2] = 0.$$

The *correlation* between two random variables X, Y is defined to be $\rho := \text{cov}(X,Y)/(\sigma_X \sigma_Y)$ for standard deviations σ_X, σ_Y. Thus it follows that independence $\implies$ zero covariance $\implies$ uncorrelatedness.

While X_1, X_2 being uncorrelated does not imply independence in general, remarkably, jointly Gaussian random variables are independent if and only if they are uncorrelated! Let's see why this holds.

Theorem 1. *Jointly Gaussian random variables are independent if and only if they are uncorrelated.*

Proof. Without loss of generality, we will consider the case of two jointly Gaussian random variables. Extensions to higher dimensions follow by the same reasoning. Suppose that X_1, X_2 are uncorrelated. Recall that the entries of the covariance matrix are $\Sigma_{i,j} = \text{cov}(X_i, X_j)$, which means that

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \text{ and } \Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix}.$$

Substituting Σ^{-1} above into the joint PDF, we find that

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right)$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right) \\
&= f_{\mathbf{X}_1}(x_1) f_{\mathbf{X}_2}(x_2).
\end{aligned}
$$

□

Note. We have shown that for *jointly Gaussian* random variables, the variables being uncorrelated implies that they are independent. This does not, however, mean that any two uncorrelated marginally normally distributed random variables are necessarily independent. To see why the variables being *jointly Gaussian* is so crucial, we will consider an example.

Example 1. Consider $X \sim \mathcal{N}(0,1)$, and $Y = WX$, where

$$
W = \begin{cases} 1 & \text{w.p. } 0.5 \\ -1 & \text{w.p. } 0.5 \end{cases}
$$

is independent of X. Notice that X and Y are uncorrelated:

$$
\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X^2 W] - 0 = \mathbb{E}[X^2]\,\mathbb{E}[W] = 0.
$$

However, they are not independent:

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

Therefore, one must ensure that the random variables are jointly Gaussian before assuming that any of these properties necessarily hold.

2.2 Linear Combinations of JG RVs are JG

Theorem 2. *Linear combinations of jointly Gaussian random variables are jointly Gaussian.*

Proof. Again, without loss of generality, we will consider the case of two jointly Gaussian random variables. Extensions to higher dimensions follow by the same reasoning. Let X_1, X_2 be jointly Gaussian. By definition 1, there exist $\boldsymbol{\mu} \in \mathbb{R}^2$ and $A \in \mathbb{R}^{2\times\ell}$ such that

$$
\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = A\mathbf{Z} + \boldsymbol{\mu} = \begin{bmatrix} A_1^\top \mathbf{Z} + \mu_1 \\ A_2^\top \mathbf{Z} + \mu_2 \end{bmatrix},
$$

where A_i is the ith row vector of A. Now, for any $C, D, c, d \in \mathbb{R}$, let $U = CX_1 + c$ and $V = DX_1 + d$. Substituting in the above expressions, we find

$$U = C(A_1^\top \mathbf{Z} + \mu_1) + c$$
$$V = D(A_2^\top \mathbf{Z} + \mu_2) + d$$
$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} CA_1 \\ DA_2 \end{bmatrix} \boldsymbol{Z} + \begin{bmatrix} C\mu_1 + c \\ D\mu_2 + d \end{bmatrix}.$$

Since U, V satisfy the form in definition 1, they are jointly Gaussian. □

2.3 MMSE and LLSE Are Equivalent

Recall from our Hilbert Space note that the minimum mean squared error estimator (MMSE) finds the function φ that minimizes $\mathbb{E}[(Y - \varphi(X))^2]$. In contrast, the linear least squares estimator (LLSE) limits φ to linear functions, finding $a, b \in \mathbb{R}$ to minimize $\mathbb{E}[(Y - a - bX)^2]$. We will now show that for jointly Gaussian random variables, the function that minimizes the mean squared error is linear.

Theorem 3. *For jointly Gaussian random variables, the MMSE $\mathbb{E}[X \mid Y]$ is equivalent to the LLSE $\mathbb{L}[X \mid Y]$.*

Proof. Let X, Y be jointly Gaussian random variables. We will first show that $X - \mathbb{L}[X \mid Y]$ and Y are uncorrelated. In the Hilbert Space note, we discussed how $X - \mathbb{L}[X \mid Y]$ is orthogonal to Y by the projection property of LLSE. Since Y and $X - \mathbb{L}[X \mid Y]$ are orthogonal, $\mathbb{E}[Y(X - \mathbb{L}[X \mid Y])] = 0$. Recall from the definition of LLSE that

$$\mathbb{L}[X \mid Y] = \mathbb{E}[X] + \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(Y)}(Y - \mathbb{E}[Y]).$$

From the linearity of expectation, it follows that

$$\mathbb{E}[X - \mathbb{L}[X \mid Y]] = \mathbb{E}\left[X - \mathbb{E}[X] - \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(Y)}(Y - \mathbb{E}[Y])\right] = 0.$$

This means that

$$\begin{aligned} \operatorname{cov}(Y, X - \mathbb{L}[X \mid Y]) &= \mathbb{E}[Y(X - \mathbb{L}[X \mid Y])] - \mathbb{E}[Y]\,\mathbb{E}[(X - \mathbb{L}[X \mid Y])] \\ &= 0 - \mathbb{E}[Y] \cdot 0 = 0. \end{aligned}$$

Therefore, $X - \mathbb{L}[X \mid Y]$ and Y are uncorrelated. By Theorem 2, $X - \mathbb{L}[X \mid Y]$ and Y are jointly Gaussian since they are linear combinations of X and Y. Thus, by Theorem 1, the uncorrelated jointly Gaussian $X - \mathbb{L}[X \mid Y]$ and Y must be independent.

We know that functions of independent random variables are independent (see Lemma 1 in the Appendix). This implies that $X - \mathbb{L}[X \mid Y]$ and $\varphi(Y)$ are independent for all functions $\varphi(\cdot)$. Independent random variables are uncorrelated, so

$$\begin{aligned}\operatorname{cov}(\varphi(Y), X - \mathbb{L}[X \mid Y]) &= \mathbb{E}[\varphi(Y)(X - \mathbb{L}[X \mid Y])] - \mathbb{E}[\varphi(Y)] \cdot 0 \\ &= \mathbb{E}[\varphi(Y)(X - \mathbb{L}[X \mid Y])] = 0.\end{aligned}$$

Therefore $X - \mathbb{L}[X \mid Y]$ is orthogonal to $\varphi(Y)$ for every $\varphi(\cdot)$. By the orthogonality property of the MMSE, $\mathbb{L}[X \mid Y] = \mathbb{E}[X|Y]$. □

3 [Optional] Covariance Matrices

3.1 Positive Semidefiniteness

In general, covariance matrices are *positive semidefinite* (PSD).

Definition 3. A symmetric matrix M is PSD if the following equivalent conditions hold:

1. $M = AA^\top$ for some matrix A.

2. For all vectors x, $x^\top M x \geq 0$.

3. M has all real, nonnegative eigenvalues.

Clearly the first point is true for covariance matrix of jointly Gaussian random variables by definition. In the following subsections, we shall see how to interpret each of these statements in different ways.

Note. In order for the PDF of a multivariate normal to be defined, the covariance matrix must be positive definite, meaning that for all x, $x^\top \Sigma x > 0$ or that Σ has all real positive eigenvalues.

3.2 Projection

Suppose we had a jointly Gaussian vector $\boldsymbol{X}$ and its centered version $\hat{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{\mu}$, and wanted to find the variance when projecting $\hat{\boldsymbol{X}}$ along a particular unit direction u. By the definition of projection, this quantity is

$$\begin{aligned}
\text{var}(u^\top \hat{\boldsymbol{X}}) &= \text{var}(u^\top A\mathbf{Z}) \\
&= \text{cov}((A^\top u)^\top \mathbf{Z}, (A^\top u)^\top \mathbf{Z}) \\
&= \text{cov}\left(\sum_{i=1}^{\ell}(A^\top u)_i Z_i, \sum_{i=1}^{\ell}(A^\top u)_i Z_i\right) \\
&= \sum_{i=1}^{\ell}(A^\top u)_i^2 \,\text{cov}(Z_i, Z_i) \\
&= (A^\top u)^\top (A^\top u) \\
&= u^\top \Sigma u.
\end{aligned}$$

Thus, we can interpret the quantity $u^\top \Sigma u$ as the variance of the projection of $\hat{\boldsymbol{X}}$ along u, which must be nonnegative! Therefore the second property holds for JG random variables. (Although here we restrict ourselves to u with unit length, we can easily generalize by scaling u by a constant factor.)

3.3 Deriving the Square Root A

Suppose we are given $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and want to find an appropriate matrix A such that $\Sigma = AA^\top$. How can we do so? Well, the Spectral Theorem states that any symmetric matrix M can be decomposed as

$$M = U\Lambda U^\top,$$

where U is orthonormal and Λ is diagonal. U and Λ will also contain the eigenvector and eigenvalue pairs of M.

If M is PSD, as is the covariance matrix Σ, then the entries of Λ will be nonnegative with square root $\Lambda^{1/2}$, namely Λ with each of its diagonal entries square rooted.

With all of this, we can find one such A that works, namely $A = U\Lambda^{1/2}U^\top$. (Note that A is not unique, as $A = \Lambda^{1/2}U^\top$ also satisfies $\Sigma = AA^\top$.)

3.4 Density Level Curves

If we examine the PDF of a JG RV (assuming it has positive definite Σ, so an inverse exists), the significant term is

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

The level curves of g are the points which have equal density in the PDF. It turns out that the level curves of g are hyperellipsoids centered at $\boldsymbol{\mu}$. For additional details, reference 4.2 in the Appendix.

4 Appendix

4.1 Functions of Independent RVs Are Independent

Lemma 1 (Functions of independent RVs are independent)**.** Let X, Y be two independent random variables and g, h be real valued functions defined on the codomains of X and Y respectively. Then, $g(X)$ and $h(Y)$ are independent random variables.

Proof.

$$\begin{aligned} \mathbb{P}(g(X) \in A, h(Y) \in B) &= \mathbb{P}(X \in g^{-1}(A), Y \in h^{-1}(B)) \\ &= \mathbb{P}(X \in g^{-1}(A)) \cdot \mathbb{P}(Y \in h^{-1}(B)) \\ &= \mathbb{P}(g(X) \in A) \cdot \mathbb{P}(h(Y) \in B). \end{aligned}$$

□

4.2 Density Level Curves Continued

To get a geometric understanding, we shall work our way up in difficulty with examples. For now, let us assume the random variables are zero-mean, so we do not have to worry about the $\boldsymbol{\mu}$ term.

4.2.1 When $\Sigma = I$

Let us start by considering the level curves of g when $\Sigma = I$:

$$g(x) = x^\top \Sigma^{-1} x = x^\top x = \|x\|_2^2.$$

From this, we can clearly see that the level curves of g are hyperspheres centered at the origin.

4.2.2 When $\Sigma = \Lambda$

Things get slightly more complicated when we generalize to a positive diagonal matrix for Σ, but not by much:

$$g(x) = x^\top \Sigma^{-1} x = x^\top \Lambda^{-1} x = \sum_{i=1}^{\ell} \frac{1}{\lambda_i} x_i^2.$$

The level curves of g are now no longer hyperspheres, but hyperellipsoids! These are generalizations of ellipses to higher dimensions, and their axes are parallel to the coordinate axes. In particular, the semi-axis length in the ith coordinate direction is $\sqrt{\lambda_i}$.

4.2.3 When $\Sigma = U\Lambda U^\top$

Now let us consider the most general case:

$$g(x) = x^\top \Sigma^{-1} x = x^\top U \Lambda^{-1} U^\top x = \sum_{i=1}^{\ell} \frac{1}{\lambda_i} (U^\top x)_i^2.$$

The level curves are again hyperellipsoids with the same semi-axis lengths of $\sqrt{\lambda_i}$. However, this time, the semi-axis directions are not along the coordinate directions, but along the directions defined by the columns of U!

4.2.4 Nonzero μ

Previously we have assumed $\boldsymbol{\mu} = 0$, but what if that isn't actually true? When our random vector has nonzero mean, we effectively have a translation. The level curves of g will still remain the same shape, but will simply be moved in space such that the center is at $\boldsymbol{\mu}$ instead of the origin.