Multivariate Gaussian

We say that a random variable $X \in \mathbb{R}^D$ is a **Gaussian random vector** if there exists a vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and a symmetric positive definite matrix $\boldsymbol{R}$ such that its density can be written as

$$f_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det(\boldsymbol{R})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

The vector $\boldsymbol{\mu}$ is the mean of this distribution, and $\boldsymbol{R}$ is the covariance:

$$\boldsymbol{\mu} = \mathrm{E}[X], \quad \boldsymbol{R} = \mathrm{E}[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^{\mathrm{T}}].$$

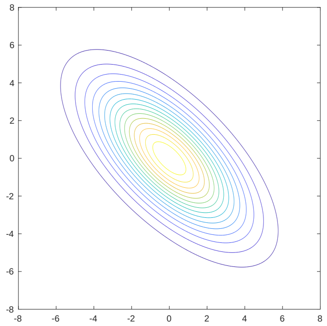
We will denote this as

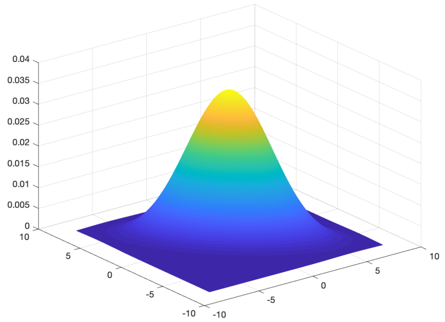
$$X \;\sim\; \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{R}).$$

The geometry of the density reflects the eigenstructure of $\boldsymbol{R}$ — the level-surfaces of the density are ellipsoids with the eigenvectors of $\boldsymbol{R}$ as axes, and radii are proportional to the eigenvalues of $\boldsymbol{R}$.

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

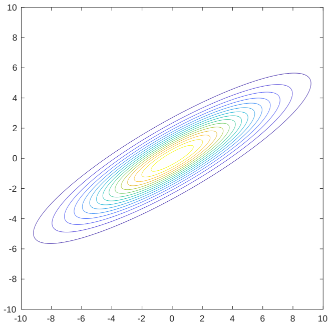
$$\boldsymbol{R} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$$

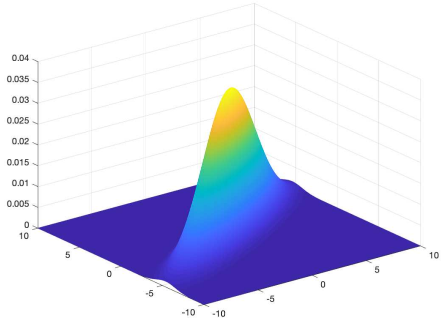




$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{R} &= \begin{bmatrix} 15.25 & 8.23 \\ 8.23 & 5.75 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \end{aligned}$$





It is clear from its definition that $\boldsymbol{R}$ is symmetric. It is also postive semidefinite[1], as

$$\begin{aligned} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}}\,\mathrm{E}[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^{\mathrm{T}}]\boldsymbol{w} &= \mathrm{E}[\boldsymbol{w}^{\mathrm{T}}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{w}] \\ = \mathrm{E}[|\boldsymbol{w}^{\mathrm{T}}(X-\boldsymbol{\mu})|^2]. \end{aligned}$$

Since $|\boldsymbol{w}^{\mathrm{T}}(X-\boldsymbol{\mu})|^2$ is always non-negative, its expectation is non-negative as well. So all the eigenvalues are $\lambda_d \geq 0$.

Other facts about the multivariate Gaussian:

1. If $X \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$, then for and $\boldsymbol{w} \in \mathbb{R}^D$, $Y = \boldsymbol{w}^{\mathrm{T}}X$ is a Gaussian (scalar) random variable with

$$Y \sim \mathrm{Normal}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{w}).$$

   You will prove this on the homework. Note in particular that this means each entry in X is itself a Gaussian random variable, as taking $\boldsymbol{w}$ as a unit vector with entry i equal to 1 and zero elsewhere, we have

$$X_i \sim \mathrm{Normal}(\mu_i, R_{ii})$$

[1] This is not just a property of the multivariate Gaussian; the covariance matrix of any random vector will be symmetric positive semidefinite.

2. If $X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$ and $\boldsymbol{A}$ is a $M \times D$ matrix, then $Y = \boldsymbol{A}X$ is a Gaussian random vector:

$$Y \sim \text{Normal}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{R}\boldsymbol{A}^{\mathrm{T}}).$$

   That the mean is $\boldsymbol{A}\boldsymbol{\mu}$ follows directly from the linearity of the expectation operator, and that each entry of Y is Gaussian follows directly from our first fact. The only thing we have to check is the expression for the covariance:

$$\begin{aligned} \mathrm{E}[(Y - \boldsymbol{A}\boldsymbol{\mu})(Y - \boldsymbol{A}\boldsymbol{\mu})^{\mathrm{T}}] &= \mathrm{E}[\boldsymbol{A}(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}] \\ &= \boldsymbol{A}\,\mathrm{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^{\mathrm{T}}]\boldsymbol{A}^{\mathrm{T}} \\ &= \boldsymbol{A}\boldsymbol{R}\boldsymbol{A}^{\mathrm{T}}. \end{aligned}$$

3. If $R_{ij} = 0$, then entries X_i and X_j are independent. You will prove this on the homework.

4. For $X \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$, let

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$$

   be the eigenvalue decomposition of $\boldsymbol{R}$. Set

$$\boldsymbol{Z} = \boldsymbol{V}^{\mathrm{T}}X.$$

   Then the covariance of Z

$$\mathrm{E}[(Z - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu})(Z - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu})^{\mathrm{T}}] = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{V} = \boldsymbol{\Lambda},$$

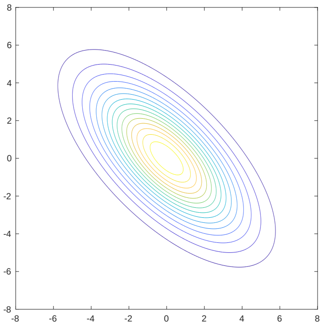
   is diagonal. Thus the entries of Z are independent, with variances equal to the eigenvalues:

$$\begin{aligned} Z_1 &\sim \text{Normal}(\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{\mu}, \lambda_1) \\ Z_2 &\sim \text{Normal}(\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{\mu}, \lambda_2) \\ &\vdots \\ Z_D &\sim \text{Normal}(\boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{\mu}, \lambda_D). \end{aligned}$$

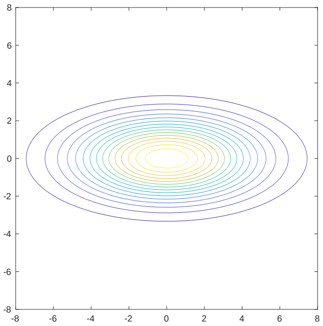
So transforming into the $\boldsymbol{V}$ domain decorrelates the entries of X. This is called the **Karhunen-Loeve** transform.

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{R} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$$



$$\downarrow \quad Z = \boldsymbol{V}^{\mathrm{T}} X$$



$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{V} = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}$$

Gaussian Estimation

What does observing part of a Gaussian random vector tell us about the part that we do not observe? That is, suppose

$$X \sim \text{Normal}(\mathbf{0}, \boldsymbol{R}),$$

and then we observe the first $1, \ldots, p$ entries of X while entries $p+1, \ldots, D$ stay hidden. We divide X into

$$X = \begin{bmatrix} X_o \\ X_h \end{bmatrix}, \quad \text{then observe} \;\; X_o = \boldsymbol{x}_o.$$

What is the conditional density for $X_h | X_o = \boldsymbol{x}_o$?

It turns out that the conditional density is also Gaussian, just with a different mean and different covariance. To see this, we partition the covariance matrix into 4 parts:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_o & \boldsymbol{R}_{oh} \\ \boldsymbol{R}_{oh}^{\mathrm{T}} & \boldsymbol{R}_h \end{bmatrix}.$$

The upper left corner contains the $p \times p$ covariance matrix for the random variables that end up being observed, the lower right corner contains the $D - p \times D - p$ covariance matrix for the unobserved random variables, and $\boldsymbol{R}_{oh}$ is the *cross-correlation* matrix, that captures the dependencies between the observed and unobserved random variables.

We can also partition the inverse covariance

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{C} \\ \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{D} \end{bmatrix}.$$

Using the Schur complement (see the Technical Details section below), we can write out these blocks of the inverse as

$$\begin{aligned}\boldsymbol{D} &= (\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh})^{-1}\\ \boldsymbol{C} &= -\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}\boldsymbol{D}\\ \boldsymbol{B} &= \text{(something)}\end{aligned}$$

We could write down the expression for $\boldsymbol{B}$ if we really wanted to, but it is long and complicated and it ends up that we don't use it. We will use these expressions later, but to ease the notation below, we will stick with $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$.

We can now compute the conditional density using

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = \frac{f_{X_o,X_h}(\boldsymbol{x}_o, \boldsymbol{x}_h)}{f_{X_o}(\boldsymbol{x}_o)}.$$

The numerator is proportional to

$$\begin{aligned}f_{X_o,X_h}(\boldsymbol{x}_o, \boldsymbol{x}_h) &\propto \exp\left(-\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_o^{\mathrm{T}} & \boldsymbol{x}_h^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{B} & \boldsymbol{C}\\ \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_o\\ \boldsymbol{x}_h\end{bmatrix}\right)\\ &= \exp\left(-\frac{1}{2}\left[\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}_o + \boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h\right]\right).\end{aligned}$$

The first term above does not depend on $\boldsymbol{x}_h$, so we can write the conditional density as

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = g(\boldsymbol{x}_0)\,\exp\left(-\frac{1}{2}\left[\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h\right]\right),$$

where $g(\boldsymbol{x}_0)$ is a function that incorporates $1/f_{X_o}(\boldsymbol{x}_o)$ and $\exp(-\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}_o/2)$ along with some constants. We are not too worried about what g actually is, just that it does not depend on $\boldsymbol{x}_h$.

To show that $X_h|X_o$ is a Gaussian random vector, we need a density that looks like

$$(\text{stuff with no } \boldsymbol{x}_h) \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}_h - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{x}_h - \boldsymbol{\mu})\right).$$

To get our conditional density in this form, we *complete the square* in the exponent. You can easily check that the following relation holds:

$$\begin{aligned}&\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x}_h + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o + \boldsymbol{x}_h^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x}_h = \\ &\qquad (\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o) - \boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o.\end{aligned}$$

Again, the last term above does not depend on $\boldsymbol{x}_h$. Thus we can write

$$f_{X_h}(\boldsymbol{x}_h|\boldsymbol{x}_0) = h(\boldsymbol{x}_o)\,\exp\left(-\frac{1}{2}(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{x}_h + \boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o)\right),$$

where $h(\boldsymbol{x}_o) = g(\boldsymbol{x}_o)\exp(\boldsymbol{x}_o^{\mathrm{T}}\boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}_o/2)$. Plugging in the expressions for $\boldsymbol{C}$ and $\boldsymbol{D}$ above, we see that $X_h|X_o = \boldsymbol{x}_o$ is a Gaussian random vector

$$X_h|X_o = \boldsymbol{x}_o \;\sim\; \mathrm{Normal}(\boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{x}_o, \boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}).$$

So given the observations $X_o = \boldsymbol{x}_o$, our best (MMSE) guess for $\boldsymbol{h}_h$ is the conditional mean:

$$\hat{\boldsymbol{x}}_h = \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{x}_o.$$

The MSE we will incur with this choice is

$$\mathrm{E}[\|\hat{\boldsymbol{x}}_h - X_h\|_2^2|X_o = \boldsymbol{x}_o] = \mathrm{trace}(\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}).$$

It is a fact that $\mathrm{trace}(\boldsymbol{R}_h - \boldsymbol{R}_{oh}^{\mathrm{T}}\boldsymbol{R}_o^{-1}\boldsymbol{R}_{oh}) \leq \mathrm{trace}(\boldsymbol{R}_h)$ (why?), so the observing $X_o = \boldsymbol{x}_o$ also reduces the mean-squared error associated with our best guess.

Notice that for zero-mean Gaussian random variables, $R[i,j] = 0$ if and only if X_i and X_j are independent. Above, this means that if X_o and X_k are independent, we will have $\boldsymbol{R}_{oh} = \boldsymbol{0}$, and the conditional distribution for X_h is no different from its original marginal (exactly as we would expect).

Example: Suppose a Gaussian random vector $X \in \mathbb{R}^2$ has

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix}.$$

We will observe $X_1 = x_1$ and see how it affects our outlook on X_2. Before the observation, we have

$$X_2 \sim \text{Normal}(0, 6),$$

so our best estimate at for X_2 is 0, and the mean-square error of this estimate is 6.

Now suppose we observe $X_1 = 4$. How does the distribution for X_2 change? Using the above ($\boldsymbol{R}_o = 6, \boldsymbol{R}_{oh} = -4, \boldsymbol{R}_h = 6$), we have

$$X_2 | X_1 = 4 \ \sim \text{Normal}\left(-\frac{8}{3}, \frac{10}{3}\right).$$

So given $X_1 = 4$, the best estimate for X_2 is now $-8/3$, and the mean-square error for that estimate is $10/3$.

Technical Details: The Schur Complement

Suppose that $\boldsymbol{M}$ is an invertible matrix broken into four blocks:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{bmatrix}.$$

If $\boldsymbol{M}_{22}$ is invertible, then the inverse of $\boldsymbol{M}$ can be expressed in terms of these blocks using the *Schur complement* of $\boldsymbol{M}$ in $\boldsymbol{M}_{22}$:

$$\boldsymbol{S} = \boldsymbol{M}_{11} - \boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}.$$

Then,

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{S}^{-1} & -\boldsymbol{S}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1} \\ -\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}\boldsymbol{S}^{-1} & \boldsymbol{M}_{22}^{-1} + \boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{21}\boldsymbol{S}^{-1}\boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1} \end{bmatrix}$$

Similarly, if $\boldsymbol{M}_{11}$ is invertible, we can do something similar with the Schur complement of $\boldsymbol{M}$ in $\boldsymbol{M}_{11}$:

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{M}_{11}^{-1} + \boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{S}^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & -\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\boldsymbol{S}^{-1} \\ -\boldsymbol{S}^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & \boldsymbol{S}^{-1} \end{bmatrix},$$

where now

$$\boldsymbol{S} = \boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}.$$

These formulas can be checked simply by multiplying out $\boldsymbol{M}^{-1}\boldsymbol{M}$ and seeing that it is $\mathbf{I}$.