Notes on the Gaussian distribution

Marta Arias, Computer Science @ UPC

These notes are meant as an introduction of the most important aspects of the multivariate Gaussian distribution. Understanding this distribution is critical in the context of machine learning since many methods relie and are built on top of this distribution. Apart from this, this distribution is notoriously important both in theory but also in practice because of the central limit theorem.

First we will see the univariate case, and then we will introduce the multivariate general case.

Much of this material is based on the excellent video series starting here. For full proofs and derivations of most of the results presented in this document, you can look at Section 2.3 of Bishop's book.

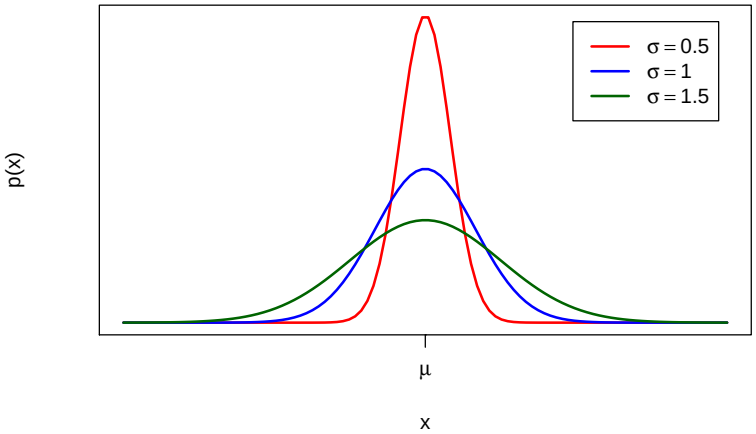
1. Univariate Gaussian distribution

Let X be a real scalar random variable. Then, we say that X is normally distributed, or X is distributed according to a Gaussian distribution with *mean* $\mu \in \mathbb{R}$ and *variance* $\sigma^2 \in \mathbb{R}^+$ (and denote this $X \sim \mathcal{N}(\mu, \sigma^2)$ if its probability density function is

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

Here, $\sqrt{2\pi\sigma^2}$ acts as a normalization constant (*independent of* x) to guarantee that $\int_x p(x; \mu, \sigma^2) = 1$. The only dependence of x in $p(x; \mu, \sigma^2)$ is through the exponent of the exponential function, $-\frac{1}{2\sigma^2}(x-\mu)^2$. Notice that this exponent is *quadratic* on x.

The following image shows different normal distributions with a fixed μ but different σ:



We observe that the pdf is symmetric around its *mean*; this can be directly observed from the fact that the exponent depends on x via $(x-\mu)^2$. Also, σ determines the *spread* of plausible values, namely its variance.

An important fact is that

- $\mathbb{E}[X] = \mu$, and

- $Var[X] = \sigma^2$.

Standard normal

A particularly famous case of Gaussian distribution is the **standard normal univariate** distribution given by $\mu = 0$ and $\sigma^2 = 1$. Often such standardized random variable is given the name Z. From the general probability density function in the univariate case, we can easily derive the pdf for $Z \sim \mathcal{N}(0, 1)$:

$$p(z; \mu = 0, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

A well-known fact is:

$\boxed{X \sim \mathcal{N}(\mu, \sigma^2) \text{ if and only if } X = Z\sigma + \mu \text{ and } Z \sim \mathcal{N}(0, 1)}$

Therefore, if we have a normal $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z \stackrel{\text{def}}{=} \frac{X-\mu}{\sigma}$ is a standard normal, that is $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. Equivalently, if we have a standard normal $Z \sim \mathcal{N}(0, 1)$, then $X \stackrel{\text{def}}{=} Z\sigma + \mu$ is distributed according to $X \sim \mathcal{N}(\mu, \sigma^2)$.

This fact gives us a way to sample from a general $X \sim \mathcal{N}(\mu, \sigma^2)$ if we have a mechanism to sample from a standard normal (and most programming languages do). All we need to do is:

1. Sample $z_1, \ldots, z_n$ from $Z \sim \mathcal{N}(0, 1)$
2. Produce samples $x_1, \ldots, x_n$ for $X \sim \mathcal{N}(\mu, \sigma^2)$ given by $x_i = \sigma z_i + \mu$ for $i = 1, \ldots, n$.

2. Multivariate Gaussian distribution

There are several equivalent definitions for the multivariate Gaussian distribution. Here, we will introduce two, and then present the pdf for the multivariate Gaussian.

Definition of the multivariate Gaussian distribution

A random vector $X = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ is said to have a multivariate Gaussian distribution if any of the following equivalent statements are true:

- Any *linear combination* $Y = \mathbf{a}^T X = a_1 X_1 + a_2 X_2 + \cdots + a_d + X_d$ with $\mathbf{a} \in \mathbb{R}^d$ is a (univariate) Gaussian.
- There exists a random vector $Z = (Z_1, \ldots, Z_m)$ with independent *standard normal* distributed Z_i, there exists a vector $\mu = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$ and a rectangular matrix $A \in \mathbb{R}^{d \times m}$ such that $X = AZ + \mu$.

We denote this with the usual $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d = (\mu_1, \ldots, \mu_d)^T$ such that $\mathbb{E}[X_i] = \mu_i$ and $\Sigma \in \mathbb{R}^{d \times d}$ is the *covariance matrix*, that is, $\Sigma_{ij} \stackrel{\text{def}}{=} Cov[X_i, X_j]$. These two parameters uniquely determine the Gaussian distribution.

By construction Σ is symmetric and positive semidefinite (PSD). In order for the pdf to be defined (to avoid degenerate cases) we also assume that Σ is **non-singular**, thus **invertible**. This makes Σ **positive definite** (PD). Throughout these notes, covariance matrices are assumed to be **positive definite**. This means that they are diagonizable, and all of their eigenvectors $\lambda_1, \ldots, \lambda_d$ are strictly positive.

In fact, we will make use of its spectral decomposition which will give us a nice geometric intuition of the distribution:

$$\Sigma = U \Lambda U^T$$

where U's columns are Σs eigenvectors and Λ is the diagonal matrix containing all of its eigenvalues. This factorization is useful for inverting it:

$$\Sigma^{-1} = U\Lambda^{-1}U^T$$

Additionally, it allows us to factorize covariance matrices in the following way:

$$\Sigma = U\Lambda U^T = U\Lambda^{1/2}\Lambda^{1/2}U^T = (U\Lambda^{1/2})(U\Lambda^{1/2})^T = AA^T$$

Namely, we can always find a matrix A such that $\Sigma = AA^T$.

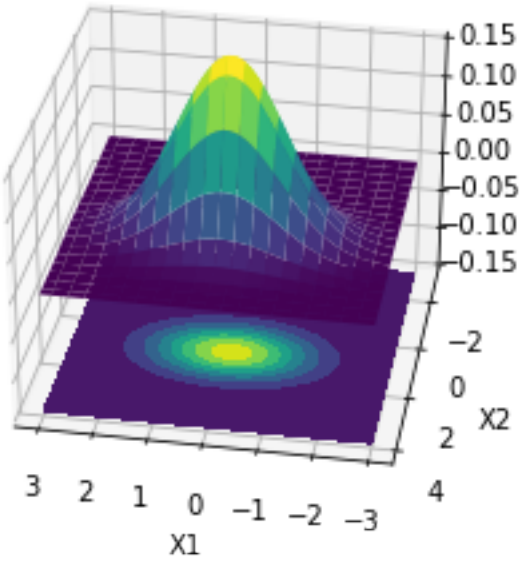
Examples of Gaussians

In order to start building some understanding and intuition for the multivariate Gaussian, let us show a way of creating multivariate Gaussians from univariate Gaussians. For this, the following result is very useful:

Fact 1. Let $X_1, \ldots, X_d$ be independent univariate Gaussians such that $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Then, $X = (X_1, \ldots, X_d) \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = (\mu_1, \ldots, \mu_k)$ and $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2) = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_d^2 \end{pmatrix}$.

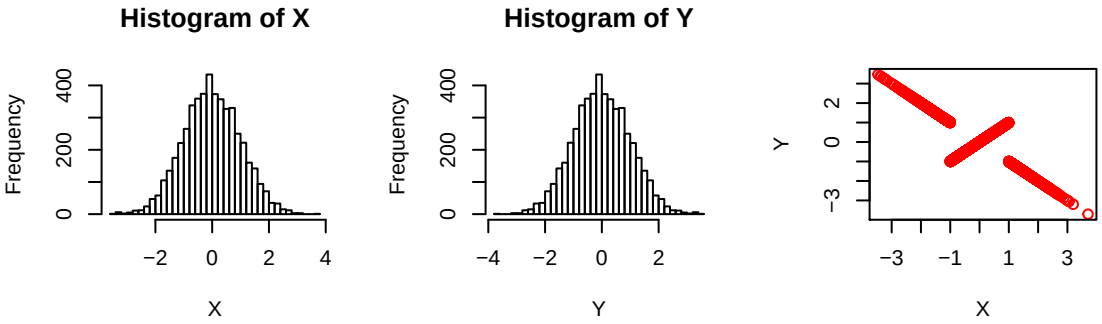
You can see an example of such a multivariate gaussian with two variables when $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(1, 0.5)$:

```
[[1.  0. ]
 [0.  0.5]]
```



Note that the joint probability being Gaussian is guaranteed by the fact that all X_i are **independent**. If this is not so, then the joint distribution may not be Gaussian, as the following counterexample shows:

Let $X \sim \mathcal{N}(0, 1)$ and define $Y = X$ if $|X| \leq 1$ and $X = -Y$ if $|X| > 1$. Clearly, X, Y are not independent however each is a univariate Gaussian. The following picture shows independent and joint distributions. Clearly, the joint distribution (red) is not Gaussian.



Finally the following is another useful fact:

Fact 2. If $X \in \mathbb{R}^d$ is Gaussian, then X_i, X_j are independent iff $Cov[X_i, X_j] = 0$.

This need not hold for non-Gaussian distributions.

Probability density function of the multivariate Gaussian

Finally we give the pdf for the multivariate Gaussian for a random variable $X \in \mathbb{R}^d$ with mean μ and (positive definite) covariance matrix Σ:

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

where $|2\pi\Sigma| = \det(2\pi\Sigma) = (2\pi)^d \det\Sigma$.

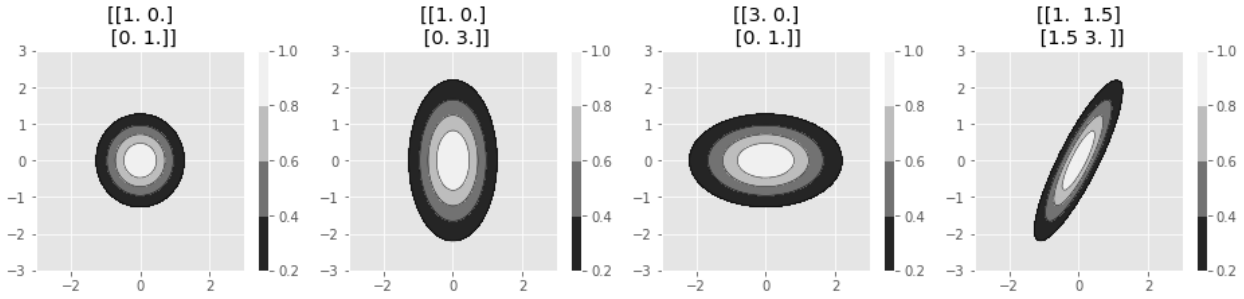
Again, the only dependence of $\mathbf{x}$ is through the exponent so $|2\pi\Sigma|$ is a normalization constant to guarantee that $\int_{\mathbf{x}} p(\mathbf{x}; \mu, \Sigma)) = 1$.

Note that the univariate case is indeed a special case where we have $\Sigma = (\sigma^2)$ and so the exponent $-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)$ becomes $-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)$ which is indeed equivalent to the expression of the univariate pdf that we gave in the beginning of this document.

The exponent is a quadratic form on $\mathbf{x}$ which turns out to be useful for visualization, since this means that (in general) when A is positive definite, solutions x s.t. $x^TAx = c$ for some constant $c \in \mathbb{R}$ form ellipsoids. This means that equal-density contours for the multivariate Gaussian are also ellipsoids, which in 2D we can visualize as ellipses. It is useful to think of the gaussian density as

$$\text{pdf} \propto \exp\{-\text{quadratic form}\}$$

Thus, we can visualize bivariate normals with contour plots of equal density given by ellipses centered at μ.



Alternative parameterization

In some cases it is prefered to use a different parameterization of the Gaussian distribution. In particular, in the context of Bayesian learning, it may be beneficial to work with **precision** instead of covariances. The **precision matrix** is simply the *inverse* of the covariance matrix. For a univariate Gaussian, the precision is thus the reciprocal of the variance.

Thus, using this alternative parameterization, we have that a univariate Gaussian is represented by *mean* μ and *precision* a, and its pdf is:

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

For the multivariate case, we use the *precision matrix* $Q = \Sigma^{-1}$[1]. Thus, the multivariate pdf is:

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

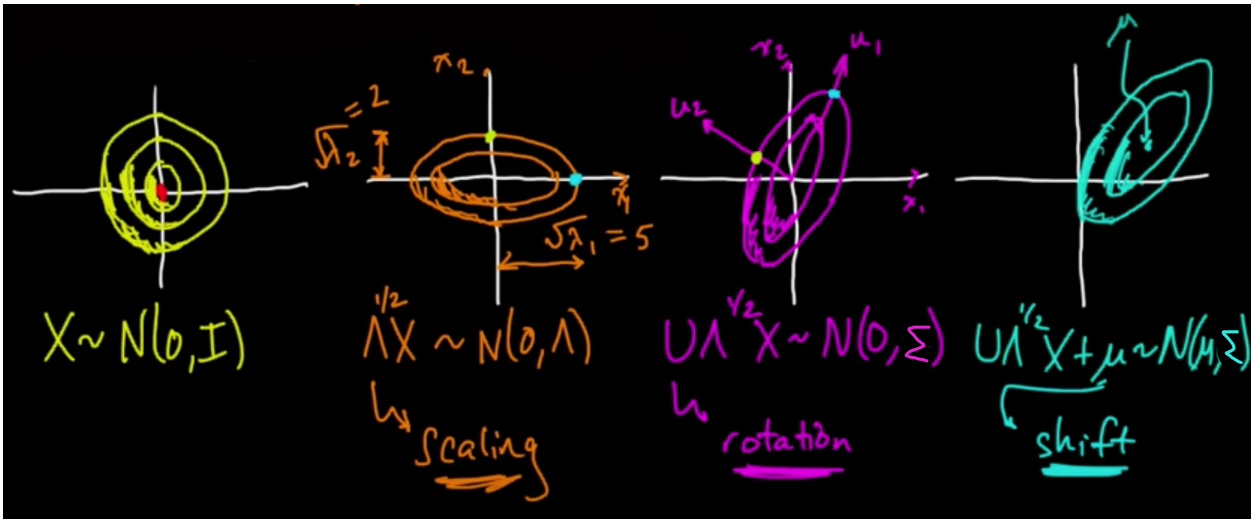
Affine property of the multivariate Gaussian

Fact 3. If $X \in \mathbb{R}^d \sim \mathcal{N}(\mu, \Sigma)$, then $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$, where $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite, $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$.

This is an extremely useful fact that allows us to construct a multivariate Gaussian distribution with mean μ and covariance Σ from univariate standard normal ones in the following way:

1. Let $X_i \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, d$ be **independent** univariate standard normals.
2. Then, $X = (X_1, \ldots, X_d) \sim \mathcal{N}(0, I)$ where $I \in \mathbb{R}^{d \times d}$ is the identity matrix. (using Fact 1)
3. Let $A = U\Lambda^{1/2}$ so that $\Sigma = AA^T$ (remember U is the matrix with eigenvectors as columns, and Λ is diagonal containing eigenvalues)
4. Finally $AX + \mu \sim \mathcal{N}(\mu, \Sigma)$ (using Fact 3)

In the particular case of a bivariate distribution we can visualize what these linear transformations are doing to the distributions we obtain step-by-step (taken from this video):

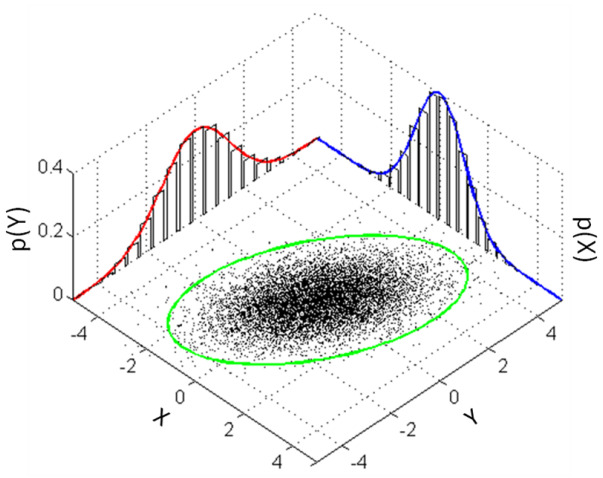


[1] It is common to use the greek letter Λ for the precision matrix however to avoid confusion with the diagonal matrix with eigenvalues that we used in the spectral decomposition we are using Q instead.

We can also perform the inverse operation and "standardize" any Gaussian in the following way:

1. Let $X \sim \mathcal{N}(\mu, \Sigma)$
2. Construct $Z = A^{-1}(X - \mu)$ so that $Z \sim \mathcal{N}(0, I)$ where $\Sigma = AA^T$ and $A^{-1} = \Lambda^{-1/2}U^T$.

Marginal distribution



Let us start with the intuition in the bivariate case:

Fact 4. If $X = (X_1, X_2)$ is a bivariate Gaussian, then the *marginals* X_1 and X_2 are also Gaussian.

Proof: Assume $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = (\mu_1, \mu_2)^T$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$. We use Fact 3 with $A = (1\ 0)$ and $b = 0$ so that $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$. Since $AX + b = X_1$, $A\mu + b = \mu_1$ and $A\Sigma A^T = \sigma_1^2$, we obtain $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$. With $A = (01)$ we obtain similarly that $X_1 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

This fact generalizes to marginalizing over any subsets of variables for d-dimensional $X = (X_1, \ldots, X_d)$.

Fact 5. Let $X = (X_1, \ldots, X_d)$ be a multivariate Gaussian with mean μ and covariance matrix Σ. We partition the variables into two sets (first p, last $q = d - p$) so that

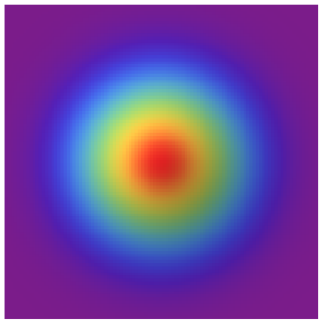
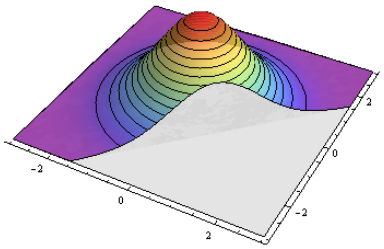
- $X = (X_a, X_b)^T$ with $X_a = (X_1, \ldots, X_p)^T$ and $X_b = (X_{p+1}, \ldots, X_k)^T$
- $\mu = (\mu_a, \mu_b)^T$ with $\mu_a = (\mu_1, \ldots, \mu_p)^T$ and $\mu_b = (\mu_{p+1}, \ldots, \mu_k)^T$
- $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$ with $\Sigma_{aa} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_p^2 \end{pmatrix}$

Then: $X_a \sim \mathcal{N}(\mu_a, \Sigma_{aa})$

To prove this, we essentially follow the same strategy but using $A = \begin{pmatrix} I & 0 \end{pmatrix}$, where $A \in \mathbb{R}^{p \times d}$ whose first p columns contain the *indentity* matrix and the rest are all 0.

Conditional distribution

Now we look at the distribution that we obtain when we *condition* on a particular value for a subset of the variables of a multivariate Gaussian. This corresponds to "slicing" through the pdf:

Fact 6. Let X, μ, Σ be like in Fact 5. Then

$$(X_a | X_b = x_b) \sim \mathcal{N}(m, S)$$

where

- $m = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$, and
- $S = C_{aa} - C_{ab}C_{bb}^{-1}C_{ba}$

This result is going to be very useful for deriving results in the context of Bayesian linear regression. We do not show it here as the proof requires a fair amount of linear algebra. For a full development of this result you can consult Section 2.3 of Bishop's book, or Section 4.3.4 of Murphy's book.

As a particular case, suppose we have two random variables X, Y that are jointly Gaussian:

$$p(x, y; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{-1/2}} \exp\left\{-\frac{1}{2}(x - \mu_X, y - \mu_Y)\Sigma^{-1}\begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}\right\}$$

with

- $\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix}$.

Then,

$$\mu_{Y|X} = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X)$$
$$\sigma_{Y|X} = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$$

Other useful properties

Fact 7. If $X \sim \mathcal{N}(\mu_X, \Sigma_X$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ are two **independent** multivariate Gaussians, then

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y).$$

To show this, let us look at the univariate case. In this case, since X, Y are independent, Fact 1 guarantees that (X, Y) is bivariate Gaussian. And now by definition any linear combination of X,Y, for example, $X + Y$ is a univariate Gaussian.

For the general case in d dimensions, we need to show that for any vector $a \in \mathbb{R}^d$ it is the case that $a^T(X + Y)$ is Gaussian. So $a^T(X + Y) = a^T X + a^T Y$ and since $a^T X$ and $a^T Y$ are univariate independent Gaussians, from the previous result so is their sum $a^T(X + Y)$.

Finally, to show that the paramaters of $X + Y$ are the ones given it is enough to verify that $\mathbb{E}[X + Y] + \mu_X + \mu_Y$ and $Cov[X + Y] = Cov[X] + Cov[Y]$ since X, Y are independent.

3. Maximum likelihood estimation of μ and Σ

Finally we give the maximum likelihood estimates of the parameters of a multivariate Gaussian given a iid data sample $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Each $\mathbf{x}_i \in \mathbb{R}^d$ is a sample from a Gaussian with unknown mean and covariance and our task is to estimate them from the observed sample D via maximization of the log-likelihood function.

So let us first write down the log-likelihood function:

$$\begin{aligned} l(\mu, \Sigma) &= \log \prod_i \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu)^T \Sigma^{-1}(\mathbf{x}_i - \mu)\right\} \\ &= \sum_i \log \frac{1}{|2\pi\Sigma|^{1/2}} - \frac{1}{2}\sum_i (\mathbf{x}_i - \mu)^T \Sigma^{-1}(\mathbf{x}_i - \mu) \\ &= \frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_i (\mathbf{x}_i - \mu)^T \Sigma^{-1}(\mathbf{x}_i - \mu) \end{aligned}$$

To derive the ML estimate for μ and Σ we need to take partial derivatives and equate to 0. To do this for μ is quite simple however to derive the estimate for Σ is more involved. It requires taking derivatives of determinants (for the left term of the log-likelihood) and traces for the quadratic form (right term of the log-likelihood). We skip this here and directly list the ML estimates for both parameters:

$$\begin{aligned} \hat{\mu}_{ML} &= \frac{1}{n}\sum_i \mathbf{x}_i \\ \hat{\Sigma}_{ML} &= \frac{1}{n}\sum_i (\mathbf{x}_i - \hat{\mu}_{ML})(\mathbf{x}_i - \hat{\mu}_{ML})^T \end{aligned}$$

4. Mahalanobis distance

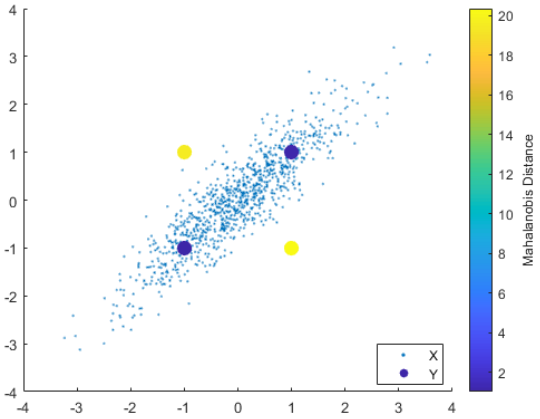
The Mahalanobis distance (or "generalized squared interpoint distance") is a dissimilarity measure between two random vectors $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$ with the same distribution with mean μ and covariance matrix Σ given by:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \Sigma^{-1}(\mathbf{x} - \mathbf{y})}$$

Notice that Σ is PSD and so Σ^{-1} has to be as well. Thus $(\mathbf{x} - \mathbf{y})^T \Sigma^{-1}(\mathbf{x} - \mathbf{y}) \geq 0$ and so $d(\mathbf{x}, \mathbf{y}) \geq 0$.

The similarity to the density for multivariate Gaussians is obvious. In fact, the Mahalanobis distance $d(\mathbf{x}, \mu)$ uniquely determines the density of $\mathbf{x}$ for a multivariate normal $\mathcal{N}(\mu, \Sigma)$. This is clear when we notice that the only dependence of the density on $\mathbf{x}$ is through the exponent $-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu) \propto d(\mathbf{x}, \mu)^2$

The Mahalanobis distance is a multi-dimensional generalization of the idea of measuring how many standard deviations away $\mathbf{x}$ is from the mean μ. But unlike Euclidean distance, it takes into account the correlations represented by the covariance matrix.



Special cases

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called a standardized Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i \frac{(x_i - y_i)^2}{\sigma_i^2}}$$

where σ_i^2 is the variance of X_i (or Y_i). Notice that each $\frac{1}{\sigma_i^2}$ is acting as a *weight* for each dimension i in the weighted sum.