Chapter 6

The Gaussian Distribution

If there is any one distribution that has traversed mathematics and found a home in cultural memory, it surely is the **Gaussian** or **normal distribution** (both names are common and we will use them interchangeably here). Not only is it super-useful in many data modelling applications, it also has a host of convenient mathematical properties, some of which we are going to explore in this chapter.

Before going into any detail, let us first motivate this distribution. What we want is a distribution on a real vector space ($\mathbb{R}^n$). We will start out with the simplest case and fist look at the Gaussian distribution on the real line. Our desiderata for the Gaussian[1] are as follows:

- The distribution should be centred around one specific point which we will call the mean
- The more distant a point is from the mean, the less probable it should be
- The distance metric should be adjustable so as to assign distant points more or less probability as needed
- Equally distant points should have the same probability, independent of their direction

6.1 The Univariate Gaussian

The Gaussian distribution is one of the most important and most widely used distributions in all of statistics. The reason is that many natural observations tend to be normally distributed. Many other distributions are

[1] Notice that Gauss' original motivation was different from ours. While we are giving a largely geometric account of the normal distribution, Gauss was concerned with finding a distribution on n independent points whose maximum likelihood estimate (See Section **??** would be the mean of those points.

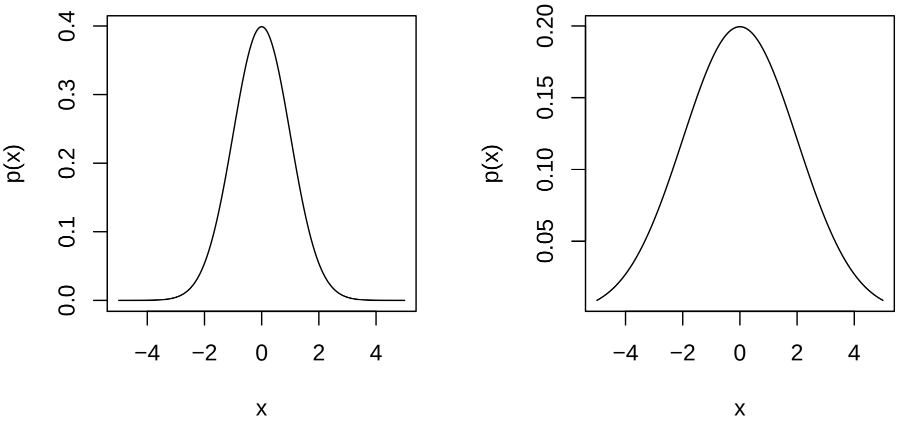

Figure 6.1: Standard normal density (left) and with variance $\sigma^2 = 2$ (right).

also based on it or can be approximated by a Gaussian. Finally, there are several mathematical properties of the Gaussian that make calculating with it rather easy. In this Section we will look at the **univariate** Gaussian distribution, that is the Gaussian distribution in one dimension. In Section 6.2 we will also see how to model data in $\mathbb{R}^n$ that is extremely complexly structured with **multivariate** Gaussian distributions.

6.1.1 Deriving the Density

What we want is a distribution that models spatial data, i.e. data that lives in some vector space. There should be a centre of mass around which the data concentrates and deviation from that centre of mass should be "penalized", meaning that the further away from the centre a data point is, the less probable it should be. Since we are interested in modelling spatial data in real vector spaces, we will choose the Euclidian distance as a distance measure. In the case of one dimension, the Euclidean distance is simply the absolute difference. For $x, y \in \mathbb{R}$, the Euclidian distance is $(y - x)^2$. Notice that this is symmetric as any good distance metric should be.

As it looks right now, all deviations are going to be penalized to the same extent. In other words: the Euclidian distance is linear in the difference of two points. What if we want to be a bit stricter and penalize points that are far away from the centre even more or conversely, if we wanted to be lenient and diminish the penalty for deviation from the centre? In such a case we would have to scale the Euclidian distance. In fact, there is a generalization of the Euclidean distance that allows for scaling. It is called the Mahalanobis distance. In the one-dimensional case, it introduces a scale factor by which the difference between two point is scaled. The Mahalanobis distance between $x, y \in \mathbb{R}$ is

$$\left(\frac{x-y}{\sigma}\right)^2$$

where $\sigma > 0$ is an adjustable scale factor. If $\sigma < 1$ it will exaggerate the difference between x and y and hence lead to a greater penalty for distant points. Conversely, if $\sigma > 1$ it will lessen the difference between x and y and therefore lead to a smaller penalty for distant points. The square of σ is called the **variance** and used to parametrize the Gaussian distribution, while σ itself is known as the **standard deviation**.

Now that we have found an appropriate (and adjustable!) distance metric, we have to turn it into a probability density. The standard way of turning any quantity into a probability density is by simply exponentiating it. This way, it is guaranteed to be positive. In the present case, we actually want that the probability decreases as the distance between the two points increases. Thus we are actually going to exponentiate the negative of the Mahalanobis distance. Finally, we might want to differentiate that distance at some point. Whenever we do so we are going to have to deal with the squaring function. In order to make our lives easier when differentiating, we also prefix the Mahalanobis distance with 1/2 before exponentiating it. The result is

$$\exp\left(-\frac{1}{2}\left(\frac{x-y}{\sigma}\right)^2\right)\,. \tag{6.1}$$

Notice that so far we have one adjustable parameter in that expression, namely the scale factor σ from the Mahalanobis distance. Initially we said that points which follow a Gaussian distribution should be arranged around a centre which is more commonly known as the **mean**. Let us call this mean μ. In order to vary the location of the centre, we turn μ into a parameter (we simply replace y with μ). To recap, μ determines the location of the centre of the Gaussian density and σ scales it. The parameters are therefore called **location parameter** and **scale parameter**, respectively. We now have an expression that is proportional to the Gaussian density. Whenever a RV X is distributed according to a normal distribution with location parameter μ and scale parameter σ, we write $X \sim \mathcal{N}(\mu, \sigma)$. The corresponding density is

$$p(x) \propto \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\,. \tag{6.2}$$

In order to get a proper density, we still need to normalize. This requires a non-trivial integration that falls without the scope of this subsection[2]. We

[2] If you are interested in seeing several different proofs, check here. Laplace's proof is probably the easiest to follow.

will just state the normalizer here. The full univariate normal density with parameters μ (mean) and σ^2 (variance) is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) . \tag{6.3}$$

Notice that we do actually never need this general density. Why? We can transform any Gaussian distribution into a **standard normal distribution**. This is the normal distribution with 0 mean and unit variance $\mathcal{N}(0,1)$. It is so important that it even has its own notation.

$$\Phi(x) = p(x) \text{ where } X \sim \mathcal{N}(0,1) \tag{6.4}$$

Exercise 6.1 *Order the points* 0.5*,* 0.8 *and* 1 *according to their probability under* $\mathcal{N}(0.7, 10)$*.*

6.2 The Multivariate Gaussian*

Our goal in this section is to define a Gaussian distribution on $\mathbb{R}^n$. This will require quite a bit of linear algebra. Readers who have not taken a linear algebra course are advised to skip this section.

Let us start out by considering a random vector whose n dimension are independent. That mean that the probability of the vector can be factorised.

$$p(\vec{x}) = \prod_{i=1}^{n} p(x_i) \tag{6.5}$$

If each of the dimensions is distributed according to the same Gaussian $\mathcal{N}(\mu, \sigma^2)$, we can easily generate random vectors of this form by making n independent from the Gaussian.

Unfortunately, this severely limits our ability to model data. Not only can we never model correlations between dimensions, we also require that all dimensions have the same variance. The data that we can model needs to be extremely homogeneous.

We could lessen this problem by drawing each random dimension from a different Gaussian. This way, we would be able to assign different means and variances to different dimensions. However, we could still not capture covariances. What we need is a single Gaussian over $\mathbb{R}^n$. This will allow us to model (potentially) dependent dimensions. Having a mean vector with different mean values per dimension is trivial. In fact, we will further assume that the means of the dimensions are independent of each other. Thus our mean vector will simply be $\vec{\mu} = \begin{bmatrix}\mu_1 & \dots & \mu_n\end{bmatrix}$. We only demand that the variances of the dimensions may be correlated. To express such correlations we need to compactly store the variances and covariances of the dimensions. To do this, we introduce **covariance matrices**.

6.2.1 Covariance Matrices

Definition 6.2 (Covariance matrix) *A $n \times n$ matrix Σ is called a covariance matrix of an n-dimensional RV X if for $0 < i, j \leq n$*

$$\Sigma_{j,i} = cov(X_j, X_i) \ .$$

The covariance matrix is has a couple of important properties which we will use when computing with it.

1. **Symmetry:** follows from the definition and the symmetry of the covariance.

2. **positive semi-definiteness:** See below.

Notice that some authors will actually define covariance matrices to be symmetric, positive semi-definite matrices. This is fine in so far as any matrix with these properties is a valid covariance matrix. When we construct models of data, we may actually simply stipulate the (co)-variances and thus build a covariance matrix.

Proof of positive semi-definiteness Recall that a $n \times n$ matrix M is positive semi-definite (PSD) if for all $\vec{z} \in \mathbb{R}^n \setminus \{0\}$ it holds that $\vec{z}^\top M \vec{z} \geq 0$. Observe that we can write a covariance matrix Σ as the expectation of an outer product.

$$\Sigma = \mathbb{E}\left[(\vec{X} - E[\vec{X}])^\top (\vec{X} - E[\vec{X}])\right] \tag{6.6}$$

For all $\vec{z} \in \mathbb{R}^n \setminus \{0\}$ we have

$$\begin{aligned}
\vec{z}\Sigma\vec{z}^\top &= \vec{z}\mathbb{E}\left[(\vec{X} - E[\vec{X}])^\top (\vec{X} - E[\vec{X}])\right]\vec{z}^\top && (6.7)\\
&= \mathbb{E}\left[\vec{z}(\vec{X} - E[\vec{X}])^\top (\vec{X} - E[\vec{X}])\vec{z}^\top\right] && (6.8)\\
&= \mathbb{E}\left[(\vec{X} - E[\vec{X}])\vec{z}^\top \vec{z}(\vec{X} - E[\vec{X}])^\top\right] && (6.9)\\
&= \mathbb{E}\left[(\vec{X} - E[\vec{X}])c(\vec{X} - E[\vec{X}])^\top\right] && (6.10)\\
&= c\mathbb{E}\left[(\vec{X} - E[\vec{X}])(\vec{X} - E[\vec{X}])^\top\right] \geq 0 && (6.11)
\end{aligned}$$

where c is some positive constant. The result essentially follows from the linearity of expectation.

The importance of being positive semi-definite may not be immediately apparent. It lies in the fact that many results are easily proven for positive semi-definite matrices. Any result that holds for positive semi-definite matrices also holds for covariance matrices. We will occasionally use this property in our proofs below.

Another important result is based solely on the symmetry of the matrix. By the spectral theorem we know that any symmetric matrix M can be factorized as

$$M = U\Lambda U^{-1} \tag{6.12}$$

where Λ is a diagonal matrix and U is orthonormal. Let us try to interpret this decomposition. The orthonormal matrix U^{-1} is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$. It effectively rotates the input. The matrix Λ then scales the each row of the input and finally the matrix U rotates the scaled input back. From the spectral theorem we know that the entries of Λ are the eigenvalues of M. Therefore, the columns of U are the corresponding eigenvectors normalized to unit length. The decomposition thus gives us an efficient way of finding the eigenvalues of M. We are now going to show that these eigenvalues are always non-negative for PSD matrices.

Lemma 6.3 (Eigenvalues of PSD matrices are non-negative) *Assume this was not the case. Let z be an eigenvalue of a positive semi-definite matrix A with negative eigenvalue λ. Then we get $z^\top Az = z^\top \lambda z = \lambda z^\top z < 0$ which contradicts the premise that A is positive semi-definite.* □

We conclude that positive semi-definite matrices (and thus covariance matrices) only have non-negative eigenvalues. This in turn implies that PSD matrices always have roots. These roots can easily be derived as

$$M^{1/2} = U\Lambda^{1/2}\Lambda^{1/2}U^{-1} = U\Lambda^{1/2}U^{-1}U\Lambda^{1/2}U^{1} \ . \tag{6.13}$$

Covariance matrices are not always used in practice. It is sometimes more convenient to use their inverse instead. That inverse, Σ^{-1}, is called a precision matrix. The names are telling: The entries in the covariance matrix measure to what extend two dimensions grow or shrink in relation to each other. The higher that value the more deviation from the mean we will observe. The entries in the precision matrix tell us how precise (i.e. how close to the mean) the distribution is. Higher precisions means that we are going to observe less deviation from the mean vector.

6.2.2 Deriving the Density

Now that we have learned about the covariance matrix, we are all set to define the multivariate Gaussian. Let us take a step back and remind ourselves of how easy it was to generate random vectors with independent means and variances. For the multivariate Gaussian, we will have to replace the mean with a mean vector (whose components are again independent[3]) and

[3] Notice that we our presentation is taking place in a frequentist setting. In Bayesian probability theory, the claim that the dimensions of the mean vector are independent may very well be false.

a covariance matrix, changing the notation from $\mathcal{N}\left(\mu, \sigma^2\right)$ to $\mathcal{N}(\vec{\mu}, \Sigma)$. As before, the parameter values are exactly equal to the mean and (co)variance of the distribution.

While we have not yet properly defined the multivariate Gaussian, we can already explore some of its properties. By simple linearity of expectation, we have for any vector $\vec{y} \in \mathbb{R}^n$ and any random Vector $X \sim \mathcal{N}(\vec{\mu}, \Sigma)$ that $\mathbb{E}[X+y] = \mathbb{E}[X] + y$ and therefore that $X + y \sim \mathcal{N}(\vec{\mu} + y, \Sigma)$. Similarly, by properties of the (co)variance and the expecation, we know that for any matrix $A \in \mathbb{R}^n$ it holds that $var(AX) = A^2 var(X) = AA\Sigma = A\Sigma A^\top$ and therefore that $AX \sim \mathcal{N}\left(A\vec{\mu}, A\Sigma A^\top\right)$.

Taken together, the fact that $AX + \vec{y} \sim \mathcal{N}\left(A\vec{\mu} + \vec{y}, A\Sigma A^\top\right)$ is called the **affine property** of the Gaussian distribution. Any affine transformation of a Gaussian RV will again yield a Gaussian RV. We will exploit this fact to define the multivariate Gaussian distribution. Recall how easy our lives would be if all variance components in the multivariate Gaussian were independent; more easy even if the variance components were also identical. Let us start from this scenario with unit variance. The **standard multivariate Gaussian** is then simply $\mathcal{N}(0, I)$ where I is the identity matrix. Clearly, the rows and columns of this matrix are orthogonal and hence independent. Moreover, only the diagonal is populated and all diagonal values are the same, meaning that we have independent and identical variances but no covariance. We can now construct an infinitely many other multivariate Gaussians with the same covariance properties by shifting the mean. Given a RV $X \sim \mathcal{N}(0, I)$ we achieve this by defining $Y = X + \vec{\mu} \sim \mathcal{N}(\vec{\mu}, I)$ (this follows from the affine property), where $\vec{\mu}$ is our desired mean.

Now that we can derive multivariate Gaussians with any mean we like, let us turn to the covariance matrix. We can change the identical variance by simply multiplying a standard normal RV with a scalar of our choice. Formally, if $X \sim \mathcal{N}(0, I)$ and $\sigma \in \mathbb{R}$ then $Y = \sigma X \sim \mathcal{N}(0, \sigma I \sigma) = \mathcal{N}\left(0, \sigma^2 I\right)$. This shouldn't come as too much of a surprise since this is how we would adjust the variance of a univariate Gaussian. However, we still cannot model covariance at this point.

Instead of multiplying $X \sim \mathcal{N}(0, I)$ with a scalar, let us use a matrix instead. In the interest of relieving all suspense, let us call this matrix $\Sigma^{1/2}$ (you see where we are getting at, aren't you?). By the affine property, we have $Y = \Sigma^{1/2} X \sim \mathcal{N}\left(0, \Sigma^{1/2} I \Sigma^{1/2^\top}\right) = \mathcal{N}(0, \Sigma)$.

What we can conclude from the above is that we can derive any multivariate Gaussian distribution from the standard normal multivariate normal simply by applying an appropriate affine transformation. Thus, all we need to do is to derive the density for the standard multivariate Gaussian. This is super-simple! The mean is 0 in all dimensions and the variances are identically and independently 1. For a vector $\vec{x} \in \mathbb{R}^n$ such that $\vec{x} \sim \mathcal{N}(0, 1)$ this means

$$p(\vec{\mu}) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \times 1} \exp\left(-\frac{1}{2}\left(\frac{x_i - 0}{1}\right)^2\right) \tag{6.14}$$
$$= \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2\right)\right) .$$

We know the density of the standard multivariate normal distribution and we know how to derive any other multivariate Gaussian from that distribution. Before we derive the general density for multivariate Gaussians, let us finally define multivariate Gaussian RVs.

Definition 6.4 (Multivariate Normal Distribution) *An n-dimensional random vector $\vec{X} \in \mathbb{R}^n$ has a multivariate normal distribution with an n-dimensional mean parameter $\vec{\mu}$ and an $n \times n$ covariance matrix Σ if it has the same distribution as $\mu + LZ$ where $LL^T = \Sigma$ and the dimensions of Z are i.i.d. according to a univariate standard normal distribution, i.e. $Z_i \sim \mathcal{N}(0,1)$ for $0 < i \leq n$.*

With this definition at hand, let us derive the general multivariate density. The problem is that in a covariance matrix the variances are not independent anymore. Thus, we cannot readily apply the factorization from Equation (6.14). The question now is whether we can substitute the covariance matrix with another matrix that where the variances are indeed independent. The spectral theorem answers this question positively. Recall that all square matrices can be decomposed according to Equation (6.12). The matrix Λ has its eigenvalues on the diagonal. Since it is congruent with the original matrix M, they both have the same eigenvalues. It is clear from Equation (6.13) that we can use $U\Lambda^{1/2} = \Sigma^{1/2}$ when applying the affine transformation of the standard normal distribution. For $X \sim \mathcal{N}(0, I), \vec{\mu} \in \mathbb{R}^n, A = U\Lambda^{1/2}U^{-1} \in \mathbb{R}^{n \times n}$ and $Y = AX + \vec{\mu}$ we exploit the fact that $Y \sim \mathcal{N}\left(\vec{\mu}, AIA^\top\right) = \mathcal{N}(\vec{\mu}, \Sigma)$. In the following, we use

M_i to denote the i^{th} row of a matrix.

(6.15)

$$p(\vec{y}) = p(A\vec{x} + \vec{\mu}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sum_{j=1}^{n} A_{ij}} \exp\left(-\frac{1}{2}\left(\frac{(A\vec{x})_i - \vec{\mu}_i}{\sum_{j=1}^{n} A_{ij}}\right)^2\right)$$

(6.16)

$$= \frac{1}{\sqrt{(2\pi)^n}\prod_{i=1}^{n}\sum_{j=1}^{n} A_{ij}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}((A\vec{x})_i - \vec{\mu}_i)^2 \sum_{j=1}^{n} A_{ij}^{-2}\right)$$

(6.17)

$$= \frac{1}{\sqrt{(2\pi)^n}\prod_{i=1}^{n}\sum_{j=1}^{n} A_{ij}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}((A\vec{x})_i - \vec{\mu}_i)\sum_{j=1}^{n} A_{ij}^{-2}\left((A\vec{x})_j - \vec{\mu}_j\right)^\top\right)$$

(6.18)

$$= \frac{1}{\sqrt{(2\pi)^n}\prod_{i=1}^{n}\sum_{j=1}^{n} A_{ij}} \exp\left(-\frac{1}{2}(A\vec{x} - \vec{\mu})\,\Sigma^{-1}\,(A\vec{x} - \vec{\mu})^\top\right)$$

(6.19)

$$= \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}} \exp\left(-\frac{1}{2}(A\vec{x} - \vec{\mu})\,\Sigma^{-1}\,(A\vec{x} - \vec{\mu})^\top\right)$$

Before we interpret this density (whose standardly given form is (6.19)) let us clarify the derivation. In order to change the indices in Equation (6.17) we have used the fact that for any $n \times n$ square matrix M and vector $\vec{x} \in \mathbb{R}^n$ we have the equality $\vec{x}^2 M = \vec{x} M \vec{x}^\top$. We then replaced A^2 with Σ. In the final line we have explicitly calculated the sum in the normalizer.

$$\sum_{j=1}^{n} A_{ij} = \sum_{j=1}^{n}\sum_{k=1}^{n} U_{ik} \sum_{l=1}^{n} \Lambda_{kl}^{1/2} U_{lj}^\top \tag{6.20}$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} U_{ik} \Lambda_{kk}^{1/2} U_{kj}^\top \tag{6.21}$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} U_{ik} U_{jk} \Lambda_{kk}^{1/2} = \Lambda_{ii}^{1/2} \tag{6.22}$$

In the above, line (6.21) follows because Λ is diagonal and the last identity in line (6.22) follows from the fact that U is orthogonal. The product in the normalizer is now a product of (square roots of) eigenvalues of a diagonal matrix, which is equal to the (square root of the) determinant of that matrix. Since congruent matrices have the same eigenvalues, this is the same as the determinant of Σ. This completes our derivation of the multivariate Gaussian density.

From the derivation, you probably already have some intuition of what's going on. Let us make this intuition more precise. A normal distribution with zero mean and covariance matrix σI for $\sigma \in \mathbb{R}$ defines a ball in which most of the probability mass lies. Being a ball, this structure is perfectly round, showing the same amount of deviance in all directions. A two-dimensional example of this is given in the upper left corner of Figure 6.2, where we show the density for the standard normal distribution on the two-dimensional plane.. Since the mean is zero, the ball is centred at the origin. If we change the mean, we are performing a shift of the balls centre away from the mean (lower left of Figure 6.2). The chosen mean there is $\vec{\mu} = \begin{bmatrix} 2 & -2 \end{bmatrix}$, while $\Sigma = I$ as in the standard normal case.

We can also stretch individual dimensions by of the ball by letting the elements on the diagonal of the covariance matrix vary independently (upper right of Figure 6.2). The covariance matrix used here is

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} .$$

In this case the ball turns into an ellipsoid that is still axis-aligned, however. Things become really interesting, though, when we use a full covariance matrix. Then we can define an ellipsoid with arbitrary orientation which contains most of the mass (lower right of Figure 6.2). Here, we have

$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 2 \end{bmatrix} .$$

It is interesting to see that because of the covariance the plane on which the Gaussian is defined has now been rotated.

How is all of this accomplished? By decomposing the covariance matrix, we have already seen that the covariance mostly depends on the eigenvalues of the covariance matrix. In fact, since scaling is done by $U\Lambda^{1/2}U^{-1}$, it is the square roots of the eigenvalues that define the spread. They are the dimension-wise standard deviations. The matrix of eigenvectors U transforms any given input into the eigenspace of Σ. Since this eigenspace has the same dimensionality as the original space and U is normal, U only performs a rotation. The matrix Λ performs the same mapping as Σ, only in eigenspace. As we have seen, this mapping is much simpler in eigenspace because Λ is diagonal. If there is no covariance, $\Sigma = \Lambda$ and thus $U = I$, meaning no rotation takes place. It is only when we introduce covariance that the orientation of the axis of our coordinate system changes (because transforming a vector in eigenspace is the same as transforming the base vectors of the vector's space into eigenvectors). The process of computing the multivariate Gaussian density can thus be broken down into 3 steps: map into eigenspace, apply the transformation given by Λ and map back into the original space.

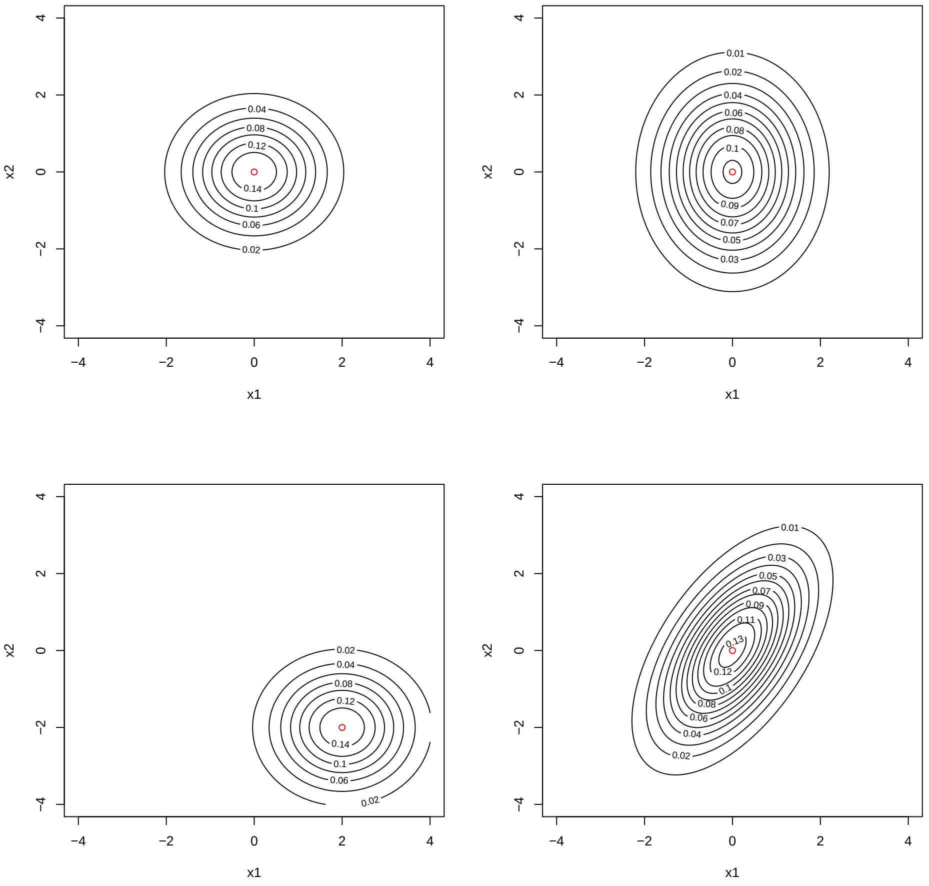

Figure 6.2: Multigauss

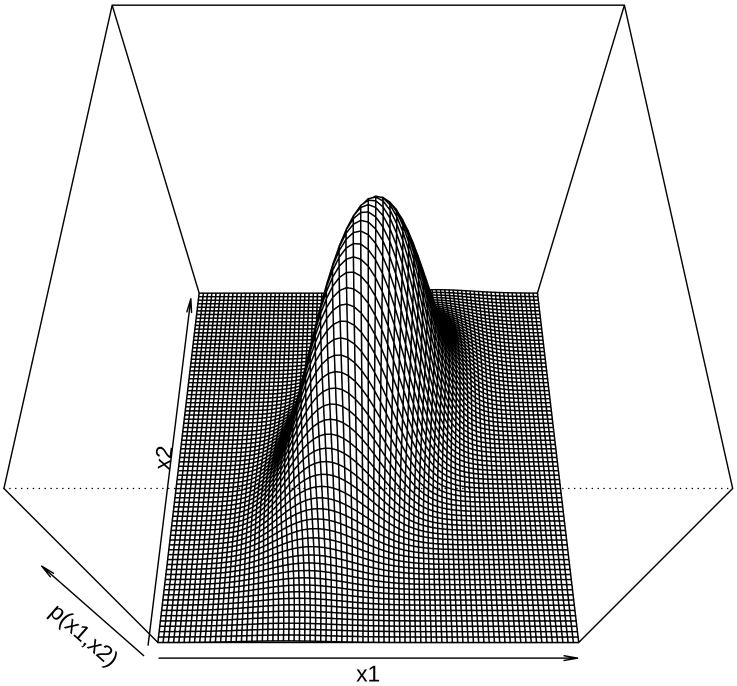

Figure 6.3: Gaussian density on top of the 2d plane. The density is the same as the one used in the lower right corner of Figure 6.2.

Notice that in Figure 6.2 we only show the plane that the two-dimensional Gaussian is defined on and annotate certain ellipses on that plane with density values. We give another visualisation in Figure 6.3 where we put the density on top of that plane.

6.2.3 Marginal and Conditional Distributions

Now that we have a thorough understanding of the multivariate normal distribution, let us look how to compute its marginal and conditional distributions.

Marginal Distribution First we re-arrange the dimensions in such a way that the k dimensions that whose marginal distribution we want to compute are the first k dimensions[4]. This can be done by appropriately swapping the rows and columns in the covariance matrix and adjusting the mean vector. Computing the marginal of the first k dimensions then comes down to projecting the n-dimensional RV $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ onto its first k dimensions. To make this explicit, we will represent the mean and covariance matrix as

$$\vec{\mu} = \begin{bmatrix} \vec{\mu}_a & \vec{\mu}_b \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \tag{6.23}$$

where part a contains the k dimensions we want to project onto and part b contains the other $n - k$ dimensions. The appropriate projection matrix is

$$P = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \tag{6.24}$$

where the identity matrix is of dimension $k \times k$. It follows that $PX = Y \sim \mathcal{N}(\vec{\mu_a}, \Sigma_{aa})$ where $Y \in \mathbb{R}^k$.

Conditional Distribution

6.2.4 Speeding Up Computation

In this subsection we will look at techniques to make computations with multivariate Gaussians more efficient.

L(ower)U(pper) Decomposition

Cholesky Decomposition

[4] This re-arrangement is not necessary but makes the exposition easier.