### 8.2.1 Multi-dimensional Gaussian distribution

In the previous of this section the Gaussian distribution has been treated with extra care. Here the theory is expanded to the $n$-dimensional case. As a first step the Gaussian distribution will be defined for an $n$-dimensional random vector. In this the density function is defined and the differntial entropy derived.

A random $n$-dimensional column vector $\boldsymbol{X} = (X_1, \dots, X_n)^T$, where $^T$ denotes the matrix transpose, is said to be Gaussian distributed if every linear combination of its entries forms a scalar Gaussian variable, i.e. if $\boldsymbol{a}^T\boldsymbol{X} = \sum_i a_i X_i \sim \mathrm{N}(\mu, \sigma)$ for every real-valued vector $\boldsymbol{a} = (a_1, \dots, a_N)^T$. Since any linaer combination of Gaussian variables is again Gaussian, the way to achieve this is to consider the case where each entrence in $\boldsymbol{X}$ is Gaussian with mean $\mu_i$ and variance $\sigma_i^2$, i.e. $X_i \in \mathrm{N}(\mu_i, \sigma_i)$. The mean of the vector $\boldsymbol{X}$ is

$$\boldsymbol{\mu} = E\big[\boldsymbol{X}\big] = (\mu_1, \dots, \mu_n)^T \tag{8.62}$$

and the covariance matrix

$$\Lambda_X = E\big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\big] = \Big(E\big[(X_i - \mu_i)(X_j - \mu_j)\big]\Big)_{i,j=1,\dots,n} \tag{8.63}$$

Clearly the diagonal elements of $\Lambda_X$ contains the variances of $\boldsymbol{X}$. The Gaussian distribution is denoted $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$.[2]

To find the density function of the distribution consider a general scaling and translation of a random variable $\boldsymbol{X}$. Let $\boldsymbol{X}$ be an $n$-dimensional random variable according to an $n$-dimensional distribution with mean $\boldsymbol{\mu}$ and covariance $\Lambda_X$. If $A$ is a square matrix of full rank and $\boldsymbol{a}$ an $n$-dimensional column vector, a new random vector $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{a}$ is formed. The mean and covariance of $\boldsymbol{Y}$ is

$$E\big[\boldsymbol{Y}\big] = E\big[A\boldsymbol{X} + \boldsymbol{a}\big] = AE\big[\boldsymbol{X}\big] + \boldsymbol{a} = A\boldsymbol{\mu} + \boldsymbol{a} \tag{8.64}$$

$$\begin{aligned}
\Lambda_Y &= E\big[(\boldsymbol{Y} - E[\boldsymbol{Y}])(\boldsymbol{Y} - E[\boldsymbol{Y}])^T\big] \\
&= E\big[A\boldsymbol{X} + \boldsymbol{a} - A\boldsymbol{\mu} - \boldsymbol{a})(A\boldsymbol{X} + \boldsymbol{a} - A\boldsymbol{\mu} - \boldsymbol{a})^T\big] \\
&= E\big[\big(A(\boldsymbol{X} - \boldsymbol{\mu})\big)\big(A(\boldsymbol{X} - \boldsymbol{\mu})\big)^T\big] \\
&= E\big[A\big(\boldsymbol{X} - \boldsymbol{\mu}\big)\big(\boldsymbol{X} - \boldsymbol{\mu}\big)^T A^T\big] \\
&= AE\big[\big(\boldsymbol{X} - \boldsymbol{\mu}\big)\big(\boldsymbol{X} - \boldsymbol{\mu}\big)^T\big]A^T = A\Lambda_X A^T
\end{aligned} \tag{8.65}$$

The idea is to transform the Gaussian vector $\boldsymbol{X}$ into a normalised Gaussian vector instead. In the case when $X$ is a one dimensional random variable, this is done with $Y = \frac{X-\mu}{\sigma}$. To see how the corresponding equation looks for the

[2]In this text it is assumed that $\Lambda_X$ has full rank. In the case it lower rank the dimensionality of the vector can be decreased.

$n$ dimensional case, some definitions and results from matrix theory is needed. For a more thorough treatment of this topic refere to e.g. [37]. Most of the results here will be given without any proofs.

Firstly, the covariance matrix is characterised to see how the square root of its inverse can be derived.

**DEFINITION 8.6** A real matrix $A$ is *symmetric*[3] if it is symmetric along the diagonal, $A^T = A$. □

If the matrix $A$ is symmetric and has an inverse, the unity matrix can be used to get $I = AA^{-1} = A^T A^{-1} = (A^{-T} A)^T = A^{-T} A$, where $^{-T}$ denotes the transpose of the inverse. Then, $A^{-1} = IA^{-1} = A^{-T} AA^{-1} = A^{-T}$. Hence, the inverse of a symmetric matrix is again symmetric. From its definition it is directly seen that the covariance matrix is symmetric, since $E[(X_i - \mu_i)(X_j - \mu_j)] = E[(X_j - \mu_j)(X_i - \mu_i)]$.

In the one-dimensional case the variance is non-negative. In matrix theory this corresponds to that the covariance matrix is positive semi-definite.

**DEFINITION 8.7** A real matrix $A$ is *positive definite* if $\boldsymbol{a}^T A \boldsymbol{a} > 0$, for all vectors $\boldsymbol{a} \neq \mathbf{0}$. □

**DEFINITION 8.8** A real matrix $A$ is *positive semi-definite*, or non-zero definite, if $\boldsymbol{a}^T A \boldsymbol{a} \geq 0$, for all vectors $\boldsymbol{a} \neq \mathbf{0}$. □

Consider the covariance matrix $\Lambda_X$ and a real valued column vector $\boldsymbol{a} \neq \mathbf{0}$. Then

$$\begin{aligned}
\boldsymbol{a}^T \Lambda_X \boldsymbol{a} &= \boldsymbol{a}^T E\big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\big]\boldsymbol{a} \\
&= E\big[\boldsymbol{a}^T (\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{a}\big] \\
&= E\big[(\boldsymbol{a}^T \boldsymbol{X} - \boldsymbol{a}^T \boldsymbol{\mu})(\boldsymbol{a}^T \boldsymbol{X} - \boldsymbol{a}^T \boldsymbol{\mu})^T)\big] = V\big[\boldsymbol{a}^T \boldsymbol{X}\big] \geq 0
\end{aligned} \tag{8.66}$$

since the variance of a one-dimensional random variable is non-negative. To conclude, the following theorem is obtained.

**THEOREM 8.6** Given an $n$-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ with mean $E\big[\boldsymbol{X}\big] = (\mu_1, \ldots, \mu_n)^T$, the covariance matrix $\Lambda_X = E\big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\big]$ is symmetric and positive semi-definite. □

[3] A complex matrix $A$ is *Hermitian* if $A^* = A$, where $^*$ denote complex conjugate and transpose. For a real matrix it is equivalent to being symmetric, i.e. $A^T = A$. In MATLAB the notation $A'$ means Hermitian transpose, $A^*$

In e.g. [37] it can be found that for every symmetric positive semi-definite matrix $A$, there exists a unique symmetric positive semi-definite matrix $A^{1/2}$ such that

$$\big(A^{1/2}\big)^2 = A \tag{8.67}$$

This matrix $A^{1/2}$ is the equivalence of the scalar square root function. Furthermore, it can be shown that the inverse of the square root is equivalent to the square root of the inverse,

$$\big(A^{1/2}\big)^{-1} = \big(A^{-1}\big)^{1/2} \tag{8.68}$$

often denoted $A^{-1/2}$. The determinant of $A^{-1/2}$ equals the inverse of the square root of the determinant,

$$\big|A^{-1/2}\big| = |A|^{-1/2} = \frac{1}{\sqrt{|A|}} \tag{8.69}$$

With this at hand, consideran an $n$-dimensional Gaussian vector, $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$. then, the normalised vector

$$\boldsymbol{Y} = \Lambda_X^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \tag{8.70}$$

has mean and covariance according to

$$E\big[\boldsymbol{Y}\big] = E\big[\Lambda_X^{-1/2}\boldsymbol{X} - \Lambda_X^{-1/2}\boldsymbol{\mu}\big] = \Lambda_X^{-1/2}E\big[\boldsymbol{X}\big] - \Lambda_X^{-1/2}\boldsymbol{\mu} = \boldsymbol{0} \tag{8.71}$$

and

$$\Lambda_Y = \Lambda_X^{-1/2}\Lambda_X\Lambda_X^{-1/2} = \Lambda_X^{-1/2}\Lambda_X^{1/2}\Lambda_X^{1/2}\Lambda_X^{-1/2} = I \tag{8.72}$$

Hence, $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{0}, I)$ is normalised Gaussian distributed with zero mean and covariance $I$. Since $\Lambda_X$ is assumed to have full rank, $|\Lambda_X| > 0$, there exists a density function that is uniquely determined by the mean and covariance. To find this, use that the entries of $\boldsymbol{Y}$ are independent and write the density function as

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_i y_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}} \tag{8.73}$$

The entropy for this vector follows from the independency as

$$H(\boldsymbol{Y}) = \sum_{i=1}^{n} H(Y_i) = n\frac{1}{2}\log(2\pi e) = \frac{1}{2}\log(2\pi e)^n \tag{8.74}$$

To calculate the entropy for the vector $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$, first consider the density function. Assume a general $n$-dimensional random vector $\boldsymbol{Z}$ with density function $f_{\boldsymbol{Z}}(\boldsymbol{z})$, and let $A$ be an $n \times n$ non-singular matrix and $\boldsymbol{a}$ an $n$ dimensional static vector. Then, form $\boldsymbol{X} = A\boldsymbol{Z} + \boldsymbol{a}$, which leads to that $\boldsymbol{Z} = A^{-1}(\boldsymbol{X} - \boldsymbol{a})$ and $d\boldsymbol{x} = |A|d\boldsymbol{z}$, where $|A|$ is the Jacobian for the variable change. Thus the density function for $\boldsymbol{X}$ can then be written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{|A|} f_{\boldsymbol{Z}}\big(A^{-1}(\boldsymbol{x} - \boldsymbol{a})\big) \tag{8.75}$$

which gives the entropy as

$$\begin{aligned}
H(\boldsymbol{X}) &= -\int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{x}) \log f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= -\int_{\mathbb{R}^n} \frac{1}{|A|} f_{\boldsymbol{Z}}\big(A^{-1}(\boldsymbol{x} - \boldsymbol{a})\big) \log \frac{1}{|A|} f_{\boldsymbol{Z}}\big(A^{-1}(\boldsymbol{x} - \boldsymbol{a})\big) d\boldsymbol{x} \\
&= -\int_{\mathbb{R}^n} f_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{1}{|A|} f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z} \\
&= -\int_{\mathbb{R}^n} f_{\boldsymbol{Z}}(\boldsymbol{z}) \log f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z} + \log |A| \int_{\mathbb{R}^n} f_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z} \\
&= H(\boldsymbol{Z}) + \log |A|
\end{aligned} \tag{8.76}$$

Hence, the following result can be stated, similar to the one-dimensional case.

**THEOREM 8.7** Let $\boldsymbol{Z}$ is an $n$-dimensional random vector with entropy $H(\boldsymbol{Z})$. If $A$ is an $n \times n$ non-singular matrix and $\boldsymbol{a}$ an $n$-dimensional static vector, then, $\boldsymbol{X} = A\boldsymbol{Z} + \boldsymbol{a}$ has the entropy

$$H(\boldsymbol{X}) = H(\boldsymbol{Z}) + \log |A| \tag{8.77}$$

□

To get back from the normalised Gaussian vector $\boldsymbol{Y}$ to $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$, use the function

$$\boldsymbol{X} = \Lambda_X^{1/2} \boldsymbol{Y} + \boldsymbol{\mu} \tag{8.78}$$

The above theorem states that the entropy for the vector $\boldsymbol{X}$ is

$$\begin{aligned}
H(\boldsymbol{X}) &= \frac{1}{2} \log(2\pi e)^n + \log |\Lambda_X|^{1/2} \\
&= \frac{1}{2} \log(2\pi e)^n |\Lambda_X| = \frac{1}{2} \log |2\pi e \Lambda_X|
\end{aligned} \tag{8.79}$$

**THEOREM 8.8** Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be an $n$-dimensional Gaussian vector with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and covariance matrix $\Lambda_X = E\big[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\big]$, i.e. $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$. Then the differential entropy of the vector is

$$H(\boldsymbol{X}) = \frac{1}{2} \log |2\pi e \Lambda_X| \tag{8.80}$$

□

An alternative way to show the above theorem is to first derive the density function for $\boldsymbol{X}$ and then use this to derive the entropy. Since this derivation will be reused later, it is also shown here. So, again use the variable change $\boldsymbol{Y} = \Lambda_X^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})$, where the Jacobian is $|\Lambda_X^{-1/2}| = \frac{1}{\sqrt{|\Lambda_X|}}$. Then

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{\sqrt{|\Lambda_X|}} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(\Lambda_X^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}))^T(\Lambda_X^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}))} \\ &= \frac{1}{\sqrt{|2\pi\Lambda_X|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Lambda_X^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \end{aligned} \tag{8.81}$$

which is the density function normally used for an $n$-dimensional Gaussian distribution.

Before progressing towards the entropy, the argument in the exponent needs some extra attension. Assume a random variable $\boldsymbol{X}$ (not necessarily Gaussian) with mean $E[\boldsymbol{X}] = \boldsymbol{\mu}$ and covariance matrix $\Lambda_X = E\big[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\big]$, and form $\boldsymbol{Y} = \Lambda_X^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})$ to get a normalised version with $E[\boldsymbol{Y}] = \boldsymbol{0}$ and $\Lambda_Y = I$. Then

$$\begin{aligned} E\big[(\boldsymbol{X}-\boldsymbol{\mu})^T\Lambda_X^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\big] &= E\big[(\boldsymbol{X}-\boldsymbol{\mu})^T\Lambda_X^{-1/2}\Lambda_X^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})\big] \\ &= E\big[\boldsymbol{Y}^T\boldsymbol{Y}\big] = E\Big[\sum_{i=1}^{n} Y_i^2\Big] = \sum_{i=1}^{n} 1 = n \end{aligned} \tag{8.82}$$

If $\boldsymbol{X}$ is Gaussian with $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$, then $Y$ is normalised Gaussian, $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{0}, I)$, and so is each of the entries, $Y_i \sim \mathrm{N}(0,1)$. Since

$$Z = (\boldsymbol{X}-\boldsymbol{\mu})^T\Lambda_X^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = \sum_{i=1}^{n} Y_i^2 \sim \chi^2(n) \tag{8.83}$$

this also gives the mean of a Chi-square distributed random variable, $E[Z] = n$.

The entropy for the Gaussian distribution can now be derived using the density function above as

$$\begin{aligned} H(\boldsymbol{X}) &= E_f\bigg[-\log \frac{1}{\sqrt{|2\pi\Lambda_X|}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T\Lambda_X^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}\bigg] \\ &= E_f\bigg[\frac{1}{2}\log|2\pi\Lambda_X| + \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T\Lambda_X^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\log e\bigg] \\ &= \frac{1}{2}\log|2\pi\Lambda_X| + \frac{1}{2}n\log e \\ &= \frac{1}{2}\log\big(e^n|2\pi\Lambda_X|\big) = \frac{1}{2}\log|2\pi e\Lambda_X| \end{aligned} \tag{8.84}$$

Looking back at Lemma 8.4 and Theorem 8.5, the corresponding result for the $n$-dimensional case can be derived. Starting with the lemma, assume that $g(\boldsymbol{x})$ is a density function for a normal distribution, $\mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$, and that $f(\boldsymbol{x})$ is an arbitrary density function with the same mean $\boldsymbol{\mu}$ and covariance matrix $\Lambda_X$. Then, the expectation of $-\log g(\boldsymbol{X})$ with respect to $g(\boldsymbol{x})$ and $f(\boldsymbol{x})$ are equal. This can be seen from the exact same derivation as above when $f(\boldsymbol{x})$ is non-Gaussian. Hence, the following lemma, corresponding to Lemma 8.4, can be stated.

**LEMMA 8.9** Let $g(\boldsymbol{x})$ be an $n$-dimensional Gaussian distribution, $\mathrm{N}(\boldsymbol{\mu}, \Lambda_X)$, with mean $\boldsymbol{\mu}$ and covariance matrix $\Lambda_X$. If $f(\boldsymbol{x})$ is an arbitrary distribution with the same mean and covariance matrix, then

$$E_f\big[-\log g(\boldsymbol{X})\big] = E_g\big[-\log g(\boldsymbol{X})\big] \tag{8.85}$$

□

To see that the Gaussian distribution maximizes the entropy consider

$$\begin{aligned} H_g(\boldsymbol{X}) - H_f(\boldsymbol{X}) &= E_g\big[-\log g(\boldsymbol{X})\big] - E_f\big[-\log f(\boldsymbol{X})\big] \\ &= E_f\big[-\log g(\boldsymbol{X})\big] - E_f\big[-\log f(\boldsymbol{X})\big] \\ &= E_f\Big[\log \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\Big] = D\big(f\,||\,g\big) \geq 0 \end{aligned} \tag{8.86}$$

**THEOREM 8.10** The $n$-dimensional Gaussian distribution maximises the differential entropy over all $n$-dimensional distributions with mean $\boldsymbol{\mu}$ and covariance matrix $\Lambda_X$. □