# Notes on Gaussian Models

compiled by D.Gueorguiev 4/4/2024

## Introductory Notes

Gaussian models are probabilistic models which involve the multivariate Gaussian (Normal) distribution usually denoted with MVN.

### Notation

*Matrices* will be denoted with uppercase bold letters, e.g. .

*Matrix entries* will be denoted by subscripted uppercase non-bold letters, e.g.

*Vectors* are assumed to be column vectors unless stated otherwise.

denotes column vector created by stacking scalars.

Similarly, will denote column vector created by concatenating vectors. Alternatively we can write:

where the notation denotes dimensional row vector formed by the scalars .

The pdf for MVN in dimensions is defined as follows:

(1)

A diagram of a red circle with arrows

Description automatically generated

Figure 1: Visualization of 2D Gaussian density. The major and minor axes of the ellipse are defined by the first two eigenvectors of the covariance matrix, namely and . See Chapter 2 of [2] for details.

The expression inside the exponent is the Mahalanobis distance between a data vector and the mean vector . We can gain a better understanding of this quantity by performing an eigendecomposition of . We have , where is an orthonormal matrix of eigenvectors satisfying , and is a diagonal matrix of eigenvalues. Using the eigendecomposition we get:

(2)

where is the -th column of , containing the -th eigenvector. Hence, we can rewrite the Mahalanobis distance as

(3)

where .

(3) tells us that the contours of equal probability density of a Gaussian lie along ellipses. This is illustrated on Figure 1. The eigenvectors determine the orientation of the ellipse, and the eigenvalues determine how elongated it is.

Note: in general, the Mahalanobis distance corresponds to Euclidean distance in a transformed coordinate system, where we shift by and rotate by .

### Applying Maximum Likelihood Estimator to Multivariate Gaussian

**Theorem** *MLE for a Gaussian*

If we have iid samples , then the MLE for the parameters is given by

(4)

(5)

That is the result of applying the MLE to multivariate Gaussian is the empirical mean and the empirical variance.

Proof:

We use the following notation and identities

– trace of a matrix

(3)

Here . We use the notation .

(4)

In order to prove (4) we use the following tensor notation:

(superscript corresponds to row vector, subscript - to column vector)

, (superscript-subscript combination on the same index implies summation ; this super-script index is called *dummy index*)

( subscripts or superscripts on the same index do not imply summation along that index )

Thus, we write

We want to prove that where (6)

Lemma: where is a function of .

We use the last Lemma in (6):

. We notice that where which proves (6).

(7)

In tensor notation (7) is rewritten as:

(8)

We will prove (8) by freezing the dummy index to a constant . Thus, the LHS of (8) becomes .

Obviously, will not be zero iff we fix and freeze . Thus .

Unfreezing and we get:

(9)

In tensor notation the LHS of (9) becomes

(10)

We use Jacobi’s formula in the form

(11)

where is the adjugate matrix of having the property

(12)

After substituting (11) and (12) into (10) we get

## Appendix

### A.1 Mahalanobis distance

Mahalanobis distance is a measure of the distance between point and a distribution . It was introduced by P.C. Mahalanobis in 1936 (see [4]).

The Mahalanobis distance is a multivariate generalization of the *standard score* : how many standard deviations away is from the mean of . This distance is zero for at the mean of and grows as moves away from the mean along each *principal component axis*. If each of these axes is rescaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance in the transformed space. Thus, the Mahalanobis distance is unitless, scale-invariant and accounts for correlations in the data set.

**Definition** *Mahalanobis distance*

Let us have a probability distribution on , generating random vectors of the form where each is a random variable. Let us denote the mean of this distribution with . Also associated with Q is the positive definite covariance matrix , where . Then the Mahalanobis distance of a point from is

.

Given two points and in , the Mahalanobis distance between them with respect to is:

Since is positive definite the inverse is also positive definite and the quadratic form expression under the square root is always positive hence the square root is defined on the set of the real numbers.

Using Whitening transformation, can be decomposed as for some real matrix, which gives us the equivalent definition:

where is the Euclidean norm. That is, the Mahalanobis distance is the Euclidean distance after applying a whitening transformation.

### A.3 Whitening Transformation

Let us define to be a column vector of random variables with non-singular covariance matrix and mean 0. Then the transformation with a whitening matrix satisfying the condition yields a whitened random vector with unit covariance. The transformation is called “*whitening*” because it changes the input vector into a *white noise* vector.

There are infinitely many possible whitening matrices that all satisfy the above condition. Commonly used choices are (*Mahalanobis/ZCA whitening*), where is the Cholesky decomposition of (*Cholesky whitening*), or eigen-system of (*PCA whitening*).

Optimal whitening transforms can be singled out by investigating the cross-covariance and cross-correlation of and . For example, the unique optimal whitening transformation achieving maximal component-wise correlation between original and whitened is produced by the whitening matrix where is the correlation matrix and the variance matrix.

### A.3 Trace of a Matrix

for details see [here](https://en.wikipedia.org/wiki/Trace_(linear_algebra)).

Let be a square matrix. Then the trace of is denoted with and is given by

Properties of trace

for all square matrices and and all scalars .

trace of a product:

//TODO: finish the paragraph on trace of a matrix

### A.4 Derivative of Matrix Logarithm

for details see [here](https://math.stackexchange.com/questions/723262/explicit-proof-of-the-derivative-of-a-matrix-logarithm).

### A.5 Jacobi’s Formula

for details see [here](https://en.wikipedia.org/wiki/Jacobi%27s_formula).

(A5.1)

Special case:

(A5.2)

Derivation:

First, let us prove the special case of the Jacobi’s formula given with (A5.2) from the general form (A5.1).

From (A5.1) it follows

(A5.3)

We have:

(A5.4)

Substituting (A5.4) in the RHS of (A5.3) leads to

from which (A5.2) follows.

Let us now derive the general form of the Jacobi’s formula (A5.1) starting with the definition of determinant of a matrix:

(A5.5)

Taking the differential of with respect to an arbitrary element of the matrix gives:

where (A5.6)

Thus, for the derivative we get

## References

[1] [Chapter 4 of Machine Learning - A Probabilistic Perspective, K. Murphy, 2012](https://github.com/dimitarpg13/probabilistic_machine_learning/blob/main/books/MachineLearningProbabilisticPerspective.pdf)

[2] [Pattern Recognition and Machine Learning, Christopher M. Bishop, 2006](https://github.com/dimitarpg13/probabilistic_machine_learning/blob/main/books/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf)

[3] [Mahalanobis distance, Wikipedia](https://en.wikipedia.org/wiki/Mahalanobis_distance)

[4] [On the Generalized Distance in Statistics, P.C. Mahalanobis, 1936](https://github.com/dimitarpg13/probabilistic_machine_learning/blob/main/articles/On_The_Generalized_Distance_In_Statistics_Mahalanobis_vol02_1936.pdf)

[5] [Introduction to Gaussian Processes and Gaussian Process Regression, Nando de Freitas, CBSC 540, UBC Jan 31, 2013](https://youtu.be/4vGiHC35j9s)

[6] [Active Learning with Gaussian Processes, Nando de Freitas, BCSC 540, UBC Feb 05, 2013](https://youtu.be/MfHKW5z-OOA)

[7] [Maximum-likelihood estimation for the multivariate normal distribution, Wiki](https://en.wikipedia.org/wiki/Estimation_of_covariance_matrices#Maximum-likelihood_estimation_for_the_multivariate_normal_distribution)