Gaussian Identities

Carl Henrik Ek

October 9, 2018

Abstract

The Gaussian distribution is often seen as having been introduced in what can be argued the first example of machine learning namely the discovery of the planet Cerces [1]. Due to its desirable mathematical properties and the central limit theorem we often encounter the Gaussian distribution in machine learning. Therefore it is a good idea to have done or at least seen the derivations of the Gaussian identities. Importantly, you do not need to know any of this by heart, that would be insanity to say the least, what you should know is have an idea of where this comes from and be able to use the results. However, if you are mathematically interested and want to understand what underpin the science of machine learning then do look at this in detail as when you are deriving new models you will most likely have to do similar calculations such as this.

It is very likely that there is errors in the following derivation, if you spot any, please let me know so that I can update the document.

1 The Gaussian

Lets first introduce the Gaussian distribution over a variable $\mathbf{x} \in \mathbb{R}^D$,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{1}$$

where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix. The mean takes the same dimensionality as the $\mathbf{x}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^D$. The characteristics of the Gaussian comes from the expression in the exponential,

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{2}$$

Let us first look at the special case where $\boldsymbol{\Sigma}$ is a diagonal matrix,

1.1 Diagonal Covariance $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & 0 & \dots & 0 \\ 0 & \Sigma_{22} & \vdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \Sigma_{DD} \end{bmatrix} \tag{3}$$

The motivation for looking at the diagonal is because the covariance matrix appears as an inverse in the exponential. The inverse of a matrix is sometimes challenging to interpret except for in the diagonal case when the inverse is reached trivially,

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\Sigma_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{\Sigma_{22}} & \vdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \frac{1}{\Sigma_{DD}} \end{bmatrix}, \tag{4}$$

by simply inverting each diagonal element.

Now when we know the inverse of the covariance lets go back and expand the exponent in the Gaussian,

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = [x_1-\mu_1,\dots,x_D-\mu_D] \begin{bmatrix} \frac{1}{\Sigma_{11}} & 0 & \dots & 0 \\ 0 & \frac{1}{\Sigma_{22}} & \vdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \frac{1}{\Sigma_{DD}} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_D-\mu_D \end{bmatrix} \tag{5}$$

$$= (x_1-\mu_1)\frac{1}{\Sigma_{11}}(x_1-\mu_1) + \dots + (x_D-\mu_D)\frac{1}{\Sigma_{DD}}(x_D-\mu_D) \tag{6}$$

$$= \sum_{i=1}^{D} \frac{1}{\Sigma_{ii}}(x_i-\mu_i)^2 \tag{7}$$

Now lets try to interpret what this actually means. First we can see that $\mathbf{x}$ only appears in a quadratic term. This means that we know that the each $(x_i-\mu_i)^2$ term will be positive. So we have a positive term multiplied by another positive term[1] which means that we have a sum of D positive terms in the exponent. Due to the minus sign in front of the exponent in Eq. 1 this means that the maximum value we will be able to get is when $x_i = \mu_i$ i.e. at the mean Figure 1. In effect the role of $\frac{1}{\Sigma_{ii}}$ is to scale the value of $(x_i-\mu_i)^2$, lets consider the following scenarios,

Σ_{ii} **is large** this means that the factor $\frac{1}{\Sigma_{ii}}$ is small and therefore we are *uncertain* about the exact value of dimension i. In other words, a large deviation from the mean have a small effect.

Σ_{ii} **is small** now a small deviation from the mean will have a large effect on the exponent, we can interpret this as we are certain about the value of this dimension.

Now lets proceed to look at the case with a general covariance structure.

[1] The definition of variance is $\mathbb{E}\left[(x-\mathbb{E}[x])^2\right]$

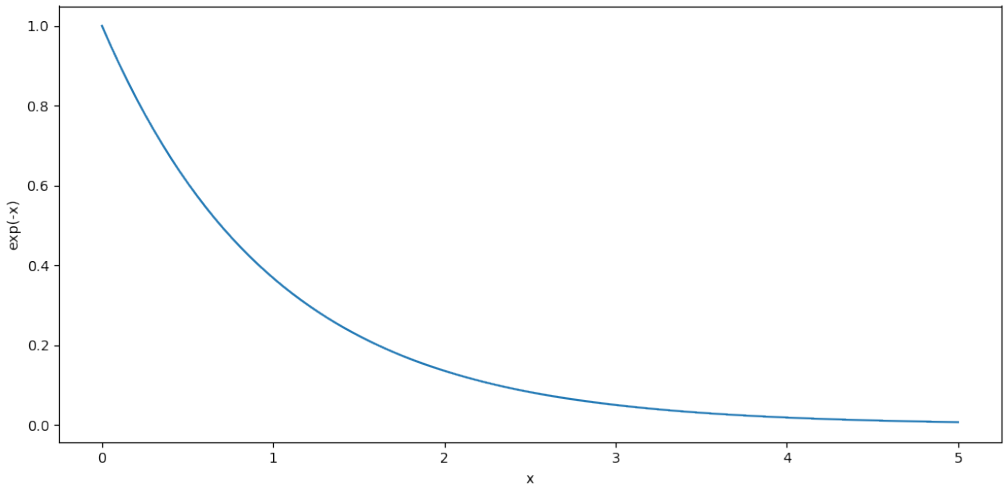

Figure 1: The above figure shows the exponential fall-off an exponential.

1.2 General Covariance

The covariance matrix specifies the covariance between each dimension, therefore the matrix is guaranteed to be square. This means that we can easily decompose and write it using its eigenvalue decomposition,

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}, \tag{8}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix and $\mathbf{U}$ is an orthonormal matrix. This decomposition allows us write the inverse covariance in a simple manner,

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{U}\boldsymbol{\Lambda}\mathbf{U})^{-1} = \mathbf{U}^{-\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}, \tag{9}$$

where we have used the fact that $\mathbf{U}$ is an orthonormal matrix i.e. $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$. Now we can write down the exponent of the Gaussian in the same way as for the diagonal case (but in a slightly less verbose manner),

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}). \tag{10}$$

Now we will split the diagonal matrix and write it as two factors as $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}$,

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) &= (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) &(11)\\ &= (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) &(12)\\ &= (\mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}))^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})) &(13)\end{aligned}$$

where we have used the rule of transposes $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$.

The new quadratic form we have derived allows for some interesting interpretations. Looking at the term that involves $\mathbf{x}$ we can see that compared to the diagonal case we pre-multiply it as,

$$\mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu}), \tag{14}$$

this is just a general linear mapping and you can think of it as a rotation of the basis to represent $\mathbf{x}$. Thinking of this as a mapping allows for further interpretation.

1.2.1 View 1

$$(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu}))^{\mathrm{T}}\mathbf{I}(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu})) \tag{15}$$

We can easily add a identity matrix in the place where we initially had the co-variance matrix. This allows us to see the mapping $\mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ as the mapping that projects the data to a space where the co-variance is identity, i.e. where each dimension is independent and have the equal variance, this is known as a *spherical* co-variance.

1.2.2 View 2

$$(\mathbf{U}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}))^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\mathbf{U}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})) \tag{16}$$

In this interpretation we keep the diagonal part of the eigendecomposition in the place of the co-variance matrix. This means that we can think of the mapping $\mathbf{U}^{\mathrm{T}}$ as the mapping that projects the data to a space where the covariance is diagonal but not necessarily diagonal.

1.3 Independent Multivariate Gaussians

Quite often we will work with independent multi-variate Gaussian variables. Say that we have a set of data $\mathbf{X} \in \mathbb{R}^{N\times D}$ so N data-points that are each D dimensional. Assuming they are independent Gaussian distributions we can write the joint probability as follows,

$$p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{N}\frac{1}{(2\pi)^{\frac{1}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})} \tag{17}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}|\boldsymbol{\Sigma}|^{\frac{N}{2}}}e^{\frac{1}{2}\mathrm{tr}((\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}))}, \tag{18}$$

where we have simply moved the product up into the exponent. Importantly the product $(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ will now generate a matrix but it is only the diagonal elements of this N matrix that corresponds to the previous expression whereby we take the `trace` operator of this matrix.

1.4 Precision Matrix

Often we will use and refer to the precision and the precision matrix of a Gaussian. The precision matrix is simply the inverse co-variance matrix. As you saw from the derivation above we usally have expressions where the co-variance appears as an inverse so it is sometimes easier to think of precision rather than variance.

$$\mathbf{\Lambda} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} & \dots & \Lambda_{1D} \\ \Lambda_{21} & \Lambda_{22} & \dots & \Lambda_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{D1} & \Lambda_{D2} & \dots & \Lambda_{DD} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1D} \\ \Sigma_{21} & \Sigma_{22} & \dots & \Sigma_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{D1} & \Sigma_{D2} & \dots & \Sigma_{DD} \end{bmatrix}^{-1} . \tag{19}$$

A large precision therefore means that we have a small variance, i.e. we are certain of the value for this dimension.

Now when we are a bit more familiar with the characteristics of the Gaussian it is time to derive the identities of the distribution so that we can use it in our learning framework.

2 Gaussian Marginal

Let us begin with the Gaussian marginal distribution, to keep things reasonably compact I will derive everything for the two dimensional case but everthing translates to more dimensions. What we want to achieve is to get from a joint Gaussian distribution $p(x_1, x_2)$ to the distribution $p(x_1)$. To simplify notation we will write our two-dimensional Gaussian using a precision matrix rather than a co-variance,

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}, \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \right) \tag{20}$$

Our task is now to integrate out x_2 from the above and reach the marginal over x_1 as,

$$p(x_1) = \int p(x_1, x_2) \mathrm{d}x_1. \tag{21}$$

The first thing we will do is to expand the exponent of the joint distribution,

$$E = -\frac{1}{2}(x_1 - \mu_1)^{\mathrm{T}} \Lambda_{11} (x_1 - \mu_1) - \frac{1}{2}(x_1 - \mu_1)^{\mathrm{T}} \Lambda_{12} (x_2 - \mu_2) \tag{22}$$

$$- \frac{1}{2}(x_2 - \mu_2)^{\mathrm{T}} \Lambda_{21} (x_1 - \mu_1) - \frac{1}{2}(x_2 - \mu_2)^{\mathrm{T}} \Lambda_{22} (x_2 - \mu_2) \tag{23}$$

$$\tag{24}$$

Being that the marginal distribution should only be a distribution over x_1 what we will now try to do is to isolate out the terms involving x_2 from the expression above. This will

be a long-windy proceedure, but its just simple algebra even though it looks like a lot.

$$E = -\frac{1}{2}((x_2^{\mathrm{T}}\Lambda_{22}x_2 - 2x_2^{\mathrm{T}}\Lambda_{22}(\mu_1 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)) \tag{25}$$

$$- 2x_1^{\mathrm{T}}\Lambda_{12}\mu_2 + 2\mu_1^{\mathrm{T}}\Lambda_{12}\mu_2 + \mu_2^{\mathrm{T}}\Sigma_{22}\mu_2 + x_1^{\mathrm{T}}\Lambda_{11}x_1 \tag{26}$$

$$- 2x_1^{\mathrm{T}}\Lambda_{11}\Sigma_{11}\mu_1 + \mu_1^{\mathrm{T}}\Lambda_{11}\mu_1) \tag{27}$$

$$= -\frac{1}{2}\left(x_2 - (\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))\right)^{\mathrm{T}}\Lambda_{22}\left(x_2 - (\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))\right) \tag{28}$$

$$+ \underbrace{\frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}x_1 - 2x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{12}\mu_1 + \mu_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mu_1\right)}_{A} \tag{29}$$

$$- \underbrace{\frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{11}x_1 - 2x_1^{\mathrm{T}}\Lambda_{11}\mu_1 + \mu_1\Sigma_{11}\mu_1\right)}_{B} \tag{30}$$

From the expansion that we have done we can see that we have three terms in the exponent. Importantly the last two terms do not include x_2 so we will now deal with them one by one. Our aim is to re-write them as quadratic expressions.

$$A = \frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}x_1 - 2x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{12}\mu_1 + \mu_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mu_1\right) \tag{31}$$

$$= \frac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})(x_1 - \mu_1)\right), \tag{32}$$

where we have used the fact that a co-variance matrix is symmetric such that $\Lambda_{12} = \Lambda_{21}$.

$$B = \frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{11}x_1 - 2x_1^{\mathrm{T}}\Lambda_{11}\mu_1 + \mu_1\Sigma_{11}\mu_1\right) = \frac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}\Lambda_{11}(x_1 - \mu_1)\right) \tag{33}$$

Importantly the two quadratic expressions we have written are both in terms of $x_1 - \mu_1$ so we can now but together the two expressions into one,

$$A - B = \frac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} - \Lambda_{11})(x_1 - \mu_1)\right). \tag{34}$$

Now lets take a step back and look at what we are aiming for. We have re-written the exponent as two separate terms, one term including x_2 and one which only includes x_1,

$$p(x_1, x_2) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}e^{E_1}e^{E_2}, \tag{35}$$

where,

$$E_1 = -\frac{1}{2}\left(x_2 - (\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))\right)^{\mathrm{T}}\Lambda_{22}\left(x_2 - (\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))\right) \tag{36}$$

$$E_2 = -\frac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})(x_1 - \mu_1).\right. \tag{37}$$

If we now go back to the formulation of the marginalisation we want to do we can exploit this structure.

$$p(x_1) = \int p(x_1, x_2)\mathrm{d}x_2 = \int \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{E_1} e^{E_2}\mathrm{d}x_2 \tag{38}$$

$$= \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{E_2} \int e^{E_1}\mathrm{d}x_2 \tag{39}$$

The above is true because the way we have re-written the exponent only E_2 depends on x_2 while E_1 depends only on x_1.

We will now proceed to integrate out x_2 from the first term in the exponent. Rather than doing this brute-force we can actually be a bit clever. If we look at E_1 we can see that it is also a quadratic form over x_2 just as the normal Gaussian,

$$E_1 = -\frac{1}{2}\left(x_2 - (\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))\right)^{\mathrm{T}} \Lambda_{22} \left(x_2 - (\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))\right), \tag{40}$$

where we can think of $(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))$ as the mean and Λ_{22} as the precision matrix. As we know that a Gaussian integrates to 1 and that the term in front of the exponential does not contain x_2 the following relationship needs to hold,

$$\int \frac{1}{(2\pi)^{\frac{D_2}{2}}|\Lambda_{22}^{-1}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_2 - \tilde{\mu}_2)^{\mathrm{T}}\Lambda_{22}(x_2 - \tilde{\mu}_2)}\mathrm{d}x_2 = 1 \tag{41}$$

$$\int e^{-\frac{1}{2}(x_2 - \tilde{\mu}_2)^{\mathrm{T}}\Lambda_{22}(x_2 - \tilde{\mu}_2)}\mathrm{d}x_2 = (2\pi)^{\frac{D_2}{2}}|\Lambda_{22}^{-1}|^{\frac{1}{2}}, \tag{42}$$

where $\tilde{\mu_2} = (\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1))$ and D_2 is the dimensionality of x_2. This means that we can re-write the expression as,

$$p(x_1) = (2\pi)^{\frac{D_2}{2}}|\Lambda_{22}^{-1}|^{\frac{1}{2}} \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{E_2} \tag{43}$$

$$= \frac{1}{(2\pi)^{\frac{D-D2}{2}}|\Lambda_{22}^{-1}|^{-\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} e^{E_2}. \tag{44}$$

Now we want to re-write the expression that involves the determinant that will be the normaliser of our distribution. To do so we will use a set of rules,

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A||D - CA^{-1}B| \tag{45}$$

$$\Rightarrow |\Sigma| = |\Sigma_{11}||\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}|\Sigma_{12}| \tag{46}$$

Now we need the final piece of the puzzle and that is we need to re-write the precision matrix Λ_{22} in terms of a co-variance term. In order to do so we will have to use what is

called a Schur complement. If you haven't seen this, or it was a long time ago I will show what they are in Section~

The Schur complement of Λ_{22} is,

$$\Lambda_{22}^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{47}$$

This means that we can simplify the terms that involves the derminants as follows,

$$|\Lambda_{22}^{-1}|^{-\frac{1}{2}}|\Sigma|^{\frac{1}{2}} = |\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|^{-\frac{1}{2}}|\Sigma_{11}|^{\frac{1}{2}}|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|^{\frac{1}{2}} \tag{48}$$

$$= |\Sigma_{11}|^{\frac{1}{2}}. \tag{49}$$

Now we can write down the full expression of the marginal distribution as follows,

$$p(x_1) = \frac{1}{(2\pi)^{\frac{D_1}{2}}|\Sigma_{11}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_1-\mu_1)^{\mathrm{T}}(\Lambda_{11}-\Lambda_{12}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}(x_1-\mu_1)} \tag{50}$$

where we have used the fact that $D = D_1 + D_2$. The final step is now to re-write the expression in precision matrices in terms of a co-variance matrix. Again we will use the Schur complement to do this,

$$\Sigma_{11}^{-1} = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}, \tag{51}$$

which leads to the final expression,

$$p(x_1) = \frac{1}{(2\pi)^{\frac{D_1}{2}}|\Sigma_{11}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_1-\mu_1)^{\mathrm{T}}\Sigma_{11}^{-1}(x_1-\mu_1)}. \tag{52}$$

The above is the marginal distribution of a Gaussian. As you can see it is actually really simple to reach it, even though the proof was rather long, you simply pick the submatrix from the mean vector and the covariance matrix that corresponds to the variables that you want and this leads to the marginal distribution.

Now when we have shown the marginal it is time to move further to look at what the conditional distribution of the Gaussian is.

3 Conditional Gaussian

Now when we have computed the marginal Gaussian distribution we will use this result to compute the conditional Gaussian distribution. To do so we will start of with the product rule,

$$p(x_1, x_2) = p(x_1|x_2)p(x_2). \tag{53}$$

As we have already proved what $p(x_2)$ is and because we define the joint distribution we already know two out of three components above. Let us start by writing up the joint distribution,

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \tag{54}$$

$$\propto e^{-\frac{1}{2}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}} \tag{55}$$

our task is now to factor out the marginal,

$$p(x_2) = \frac{1}{(2\pi)^{\frac{D_2}{2}}|\Sigma_{22}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_2-\mu_2)^{\mathrm{T}}\Sigma_{22}^{-1}(x_2-\mu_2)} \tag{56}$$

$$\propto e^{-\frac{1}{2}(x_2-\mu_2)^{\mathrm{T}}\Sigma_{22}^{-1}(x_2-\mu_2)}. \tag{57}$$

We will now use Schur complements to achieve this by expression the inverse of the full co-variance matrix decomposed in such a way that Σ_{22}^{-1} gets isolated. First lets look at the exponent of the joint distribution,

$$E = -\frac{1}{2}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \tag{58}$$

$$= -\frac{1}{2}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} I & 0 \\ \Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix}\begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix}\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \tag{59}$$

$$= -\frac{1}{2}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & -(\Sigma/\Sigma_{22})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{21}\Sigma_{22}^{-1}(\Sigma/\Sigma_{22})^{-1} & \Sigma_{22}^{-1} \end{bmatrix}^{-1}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \tag{60}$$

$$= -\frac{1}{2}(x - (\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}(x_2 - \mu_2)))^{\mathrm{T}}(\Sigma/\Sigma_{22})^{-1}(x - (\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}(x_2 - \mu_2))) \tag{61}$$

$$\underbrace{-\frac{1}{2}(x_2 - \mu_2)^{\mathrm{T}}\Sigma_{22}^{-1}(x_2 - \mu_2)}_{E_2}. \tag{62}$$

The last term in the expression above E_2 is exactly the exponent of the marginal distribution of x_2. Due to the product rule this means that we now know that the remaining term needs to be the exponent of the conditional Gaussian distribution. Therefore we only need to identify the parameters of a Gaussian to write down the posterior,

$$p(x_1|x_2) \propto e^{-\frac{1}{2}(x-\underbrace{(\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}(x_2 - \mu_2))}_{\text{mean}})^{\mathrm{T}}(\underbrace{\Sigma/\Sigma_{22}}_{\text{covariance}})^{-1}(x-(\mu_1+\Sigma_{21}\Sigma_{22}^{-1}(x_2-\mu_2)))}, \tag{63}$$

from which we get the conditional distribution as,

$$p(x_1|x_2) = \mathcal{N}(x_1|\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}), \tag{64}$$

where we have written out the Schur complement.

4 Linear Regression Posterior

Now we will look at products of Gaussians in order to formulate the posterior distribution of a Gaussian model. As we will look at linear regression as the first part in the coursework we will use this specific model as an example. We have observed input output pairs $\{y_i, x_i\}_{i=1}^N$ and which we assume have been generated from a linear model with addative Gaussian noise,

$$y_i = x_i^{\mathrm{T}} w + \epsilon \tag{65}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I). \tag{66}$$

This means that we have a likelihood function as,

$$p(y|w, x) = \mathcal{N}(y|xw, \sigma^2 I). \tag{67}$$

Our task is to reach the posterior distribution over the weights w which means we need to specify a prior distribution $p(w)$. In this case we will assume that the weights are distributed as a full covariance Gaussian with zero mean,

$$p(w) = \mathcal{N}(0, \Sigma). \tag{68}$$

In this part we will use the fact that Gaussians are self-conjugate which means we assume that we know the form of the posterior distribution which will allow us to avoid computing the evidence in Bayes rule. To start lets write up a general normal distribution and look at the exponential,

$$\mathcal{N}(x|\mu, \Sigma) \propto e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)} \tag{69}$$

$$= e^{-\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x} e^{x^{\mathrm{T}}\Sigma^{-1}\mu} e^{-\frac{1}{2}\mu^{\mathrm{T}}\Sigma^{-1}\mu}, \tag{70}$$

where we can see that the exponent contains three parts, a quadratic, a linear and a constant term with respect to the variable x. We will use this structure to identify the same parts for the posterior distribution for linear regression. Lets write up the exponent of the likelihood times the prior,

$$p(w|y, x) \propto p(y|w, x)p(w) \tag{71}$$

$$\propto -\frac{1}{2\sigma^2}(y - xw)^{\mathrm{T}}(y - xw) - \frac{1}{2}w^{\mathrm{T}}\Sigma^{-1}w \tag{72}$$

$$= \underbrace{-\frac{1}{2\sigma^2}y^{\mathrm{T}}y}_{A} + \underbrace{\frac{1}{\sigma^2}y^{\mathrm{T}}(xw)}_{B} \underbrace{-\frac{1}{2\sigma^2}(xw)^{\mathrm{T}}(xw) - \frac{1}{2}w^{\mathrm{T}}\Sigma^{-1}w}_{C}. \tag{73}$$

As we are now looking for a Gaussian distribution over w we can see that the first term A does not contain w therefore it has to be our constant term, the second term B is our linear term, while the last term C is quadratic in w. We will now use these terms in turn to identify the parameters of the posterior. First lets start with the quadratic term C.

$$C = -\frac{1}{2\sigma^2}(xw)^{\mathrm{T}}(xw) - \frac{1}{2}w^{\mathrm{T}}\Sigma^{-1}w \tag{74}$$

$$= -\frac{1}{2\sigma^2}w^{\mathrm{T}}x^{\mathrm{T}}xw - \frac{1}{2}w^{\mathrm{T}}\Sigma^{-1}w \tag{75}$$

$$= -\frac{1}{2}w^{\mathrm{T}}(\frac{1}{\sigma^2}x^{\mathrm{T}}x + \Sigma^{-1})w \tag{76}$$

As the above term is the only term that contains w in quadratic form we can now directly identify the covariance matrix of the posterior as,

$$S^{-1} = \frac{1}{\sigma^2}x^{\mathrm{T}}x + \Sigma^{-1}. \tag{77}$$

Now lets us use the linear term to identify the mean,

$$B = \frac{1}{\sigma^2}y^{\mathrm{T}}(xw) = \frac{1}{\sigma^2}w^{\mathrm{T}}x^{\mathrm{T}}y. \tag{78}$$

Now let us go back to the general exponent of a Gaussian in Eq. 4 and look at the linear term and compare this with the term B,

$$x^{\mathrm{T}}\Sigma^{-1}\mu \tag{79}$$

$$\frac{1}{\sigma^2}w^{\mathrm{T}}x^{\mathrm{T}}y. \tag{80}$$

As we have already identified the co-variance matrix we can use the linear term to solve for the mean μ,

$$w^{\mathrm{T}}S^{-1}\mu = w^{\mathrm{T}}(\frac{1}{\sigma^2}x^{\mathrm{T}}x + \Sigma^{-1})\mu = \frac{1}{\sigma^2}w^{\mathrm{T}}x^{\mathrm{T}}y \tag{81}$$

$$\Rightarrow (\frac{1}{\sigma^2}x^{\mathrm{T}}x + \Sigma^{-1})\mu = \frac{1}{\sigma^2}x^{\mathrm{T}}y \tag{82}$$

$$\Rightarrow \mu = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}x^{\mathrm{T}}x + \Sigma^{-1})^{-1}x^{\mathrm{T}}y. \tag{83}$$

Which gives the final expression for the posterior of linear regression with a Gaussian prior as,

$$p(w|y,x) = \mathcal{N}(w|\frac{1}{\sigma^2}(\frac{1}{\sigma^2}x^{\mathrm{T}}x + \Sigma^{-1})^{-1}x^{\mathrm{T}}y, (\frac{1}{\sigma^2}x^{\mathrm{T}}x + \Sigma^{-1})^{-1}) \tag{84}$$

5 Appendix

5.1 Schur Complement

Inverting matrices can be a very tedious thing and is sadly something that we have to do alot when we work with Gaussians. There is one tool though that will help us imensly and that is Schur complements. I will here derive what a Schur complement is as we will use it several times when we prove the Identities. First lets motivate the complement. If we have a block diagonal matrix the inverse of the matrix can be computed by taking the inverse of each block in turn,

$$\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}^{-1} = \begin{bmatrix} A_1^{-1} & 0 \\ 0 & A_2^{-1} \end{bmatrix}. \tag{85}$$

Now lets say that we have a general matrix M,

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix} \tag{86}$$

that we want to find the inverse of. Importantly we want to exploit the block structure of the inverse we have written above. The trick that we will do is to re-write our matrix M as a decomposition that allows us to diagonalise it into blocks. We will now try to come up with a decomposition that will "clear out" off-diagonal blocks in turn for us.

$$\begin{bmatrix} I & A \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} E + AG & F + AH \\ G & H \end{bmatrix} \tag{87}$$

Now the sub-matrix A in the expression above is for me to choose and as I want to make the product block-diagonal I will choose it such that it "clears out" the sub-matrix $F - AH$ therefore we choose it to be,

$$F + AH = 0 \tag{88}$$
$$\Rightarrow AH = -F \tag{89}$$
$$\Rightarrow AHH^{-1} = -FH^{-1} \tag{90}$$
$$\Rightarrow A = -FH^{-1}. \tag{91}$$

Now lets compute the resulting matrix after we pre-multiply[2],

$$\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} E + FH^{-1}G & 0 \\ G & H \end{bmatrix} \tag{92}$$

[2]Remember that matrix multiplication is not commutative

So we have managed to clear out the sub-matrix that is above the diagonal, next we will apply the same idea to clear out the sub-matrix below the diagonal G. In order to do this we will rather post-multiply the matrix,

$$\begin{bmatrix} E + FH^{-1}G & 0 \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} = \begin{bmatrix} E - FH^{-1}G & 0 \\ G - HB & H \end{bmatrix}. \tag{93}$$

Now we want to choose our matrix B such that the sub-matrix $G - HB$ disapears,

$$G - HB = 0 \tag{94}$$
$$\Rightarrow HB = -G \tag{95}$$
$$\Rightarrow H^{-1}HB = -H^{-1}G \tag{96}$$
$$\Rightarrow B = -H^{-1}G. \tag{97}$$

If we post-multiply with our new matrix we get the following results,

$$\begin{bmatrix} E + FH^{-1}G & 0 \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} = \begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix}, \tag{98}$$

and we have reached a block-diagonal matrix. Now let us write down the full decomposition,

$$\underbrace{\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} E & F \\ G & H \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix}}_{B} = \underbrace{\begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix}}_{C}. \tag{99}$$

This means that we have a decomposition $AMB = C$ and now we are interested in computing the inverse of M in a manner so that we exploit the block-diagonal structure.

$$(AMB)^{-1} = C^{-1} \tag{100}$$
$$B^{-1}M^{-1}A^{-1} = C^{-1} \tag{101}$$
$$\underbrace{BB^{-1}}_{=I} M^{-1}A^{-1} = BC^{-1} \tag{102}$$
$$M^{-1} \underbrace{A^{-1}A}_{=I} = BC^{-1}A \tag{103}$$
$$M^{-1} = BC^{-1}A \tag{104}$$
$$\tag{105}$$

The above means that we can write the inverse of matrix A as a product of two matrices A and B and the inverse of a block-diagonal matrix C this leads to following expression of the inverse,

$$M^{-1} = \begin{bmatrix} I & 0 \\ -HG^{-1} & I \end{bmatrix} \begin{bmatrix} (E - FH^{-1}G)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \tag{106}$$
$$= \begin{bmatrix} (E - FH^{-1}G)^{-1} & -(E - FH^{-1}G)^{-1}FH^{-1} \\ -H^{-1}G(E - FH^{-1}G)^{-1} & H^{-1} + H^{-1}G(E - FH^{-1}G)^{-1}FH^{-1} \end{bmatrix} \tag{107}$$

This means we have now expressed the inverse of the matrix M as the inverse of a block-diagonal matrix. In the case above we have blocked out the submatrix H, which sits alone in the expression of C we will therefore refer to $E - FH^{-1}G$ as the Schur complement of M with respect to H and its often indicated by M/H. If this seems unclear look at how it is used when we compute the conditional gaussian distribution and hopefully it will all come together.

6 References

References

[1] G. Foderà Serio, A. Manara, Osservatorio Astronomico Di Brera, P. Sicoli, Osservatorio Astronomico Di Sormano, Maria Ubaldo, and Nicolò Piazzi. Giuseppe piazzi and the discovery of ceres, 2002.