Course 67564 2022/23

Bayesian Machine Learning
Recitation 2: Gaussians

Prof. Yair Weiss *TA: Roy Friedman*

The distribution we will use the most in this course is the Gaussian distribution[1], also called the normal distribution.

For a single random variable x, the Gaussian distribution is defined as:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \tag{0.1}$$

Notice that this distribution can be completely described by the two parameters μ and σ - the mean and variance of the Gaussian. Because of this, we will usually write:

$$x \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{0.2}$$

to indicate that x is a Gaussian random variable, with the two parameters μ and σ; examples can be seen in Figure 1. In the same manner, we denote the PDF by:

$$p(x) = \mathcal{N}\left(x \,|\, \mu, \sigma^2\right) \qquad \left(\equiv \mathcal{N}\left(x\,;\mu,\sigma^2\right)\right) \tag{0.3}$$

The conditioning sign (or semi-colon) in $\mathcal{N}\left(x \,|\, \mu, \sigma^2\right)$ is to show that x is the variable that we are interested in, while μ and σ are the parameters that define the distribution (so, given a μ and a σ, we know the PDF of x).

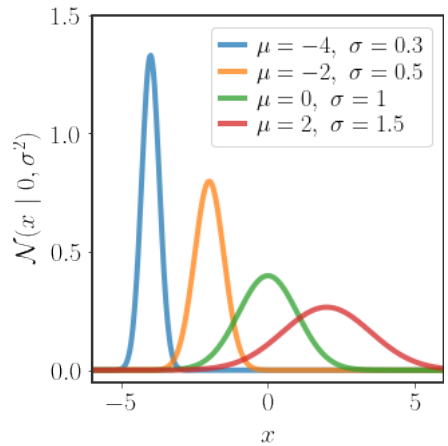

Figure 1: examples of 1D Gaussian distributions with different means (μ) and standard deviations (σ).

The multivariate version for a d-dimensional random vector $x \in \mathbb{R}^d$ is defined as:

$$\mathcal{N}(x \,|\, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right] \tag{0.4}$$

and in this case μ is also a vector and Σ is a symmetrical $n \times n$ matrix; an example can be seen in Figure 2. The term $D_M(x \,|\, \mu, \Sigma)^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$ is often called the *Mahalanobis distance* and denoted with Δ. The multivariate version for the Gaussian distribution is also called the *multivariate normal* (MVN) distribution.

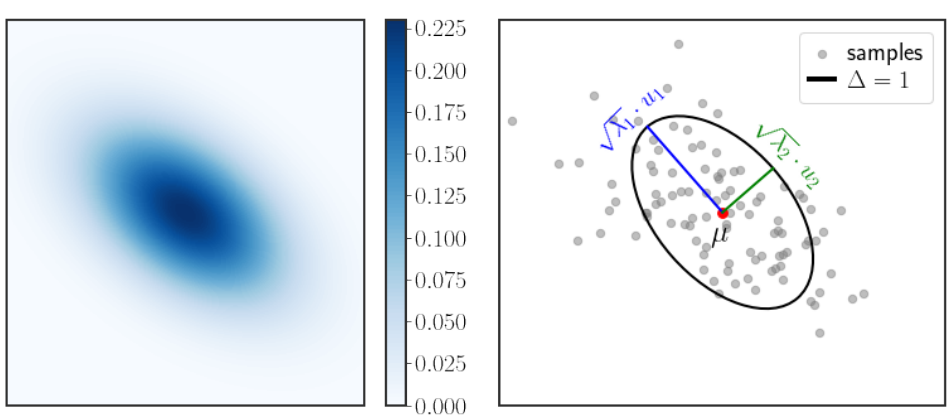

Figure 2: example of a 2D Gaussian distribution. On the left is the heatmap of the distribution - darker means higher density. On the right is the contour at $\Delta = 1$ overlayed on top of samples from the distribution. The contours of the distribution are ellipses aligned and scaled according to the eigenvectors and eigenvalues of the covariance matrix.

[1] See Bishop 2.3 for a *much* more extensive introduction to the Gaussian distribution

Meaning of μ and Σ

To understand the meaning of μ, let us look at the expectation of x under the Gaussian distribution:

$$\mathbb{E}[x] = \frac{1}{Z}\int \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right] x dx \tag{0.5}$$

$$= \frac{1}{Z}\int \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right](z+\mu)\, dz \tag{0.6}$$

where $Z = \sqrt{(2\pi)^d |\Sigma|}$ and $z = x - \mu$. We can split the integral into two - one part with z and one with μ. Since $\int \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right] dz = Z$, the second part of the integral will simply be equal to μ. The first part of the integral looks more tricky, but actually:

$$\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right] z dz = \int_{0}^{\infty} \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right] z dz + \int_{-\infty}^{0} \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right] z dz \tag{0.7}$$

$$= \int_{0}^{\infty} \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right] z dz - \int_{0}^{\infty} \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right] z dz \tag{0.8}$$

$$= 0 \tag{0.9}$$

... and we don't actually have to compute the integral! Putting the two parts together, we get:

$$\mathbb{E}[x] = \frac{1}{Z}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}z^T\Sigma^{-1}z\right] \mu dz = \mu\frac{Z}{Z} = \mu \tag{0.10}$$

so we μ is the mean of the Gaussian distribution.
In a similar manner (with many more tricks), we can show that Σ is the covariance of the Gaussian, i.e. that:

$$\mathbb{E}\left[xx^T\right] = \Sigma + \mu\mu^T$$

so that:

$$\mathrm{cov}[x] = \mathbb{E}\left[xx^T\right] - \mathbb{E}[x]\,\mathbb{E}\left[x^T\right] = \Sigma \tag{0.11}$$

1 Geometry of the Gaussian Distribution

In 1D, the Gaussian distribution takes the form of the famous bell curve in Figure 1 and is easy to view. However, in multiple dimensions it is not so clear what the geometry of the distribution actually looks like. We can gain insight by considering the EVD of the covariance matrix Σ (remember, this decomposition exists since Σ is symmetric):

$$\Sigma = UDU^T$$

where D is a diagonal matrix with the eigenvalues λ_i on the diagonal and U is an orthonormal matrix (so $UU^T = I$) with the eigenvectors u_i as it's rows, such that for all i:

$$\Sigma u_i = \lambda_i u_i \tag{1.1}$$

Recall that the eigenvectors are orthogonal to each other, and we can choose eigenvectors that are normalized, so for all $i \neq j$ we have $u_i^T u_j = 0$ and $u_i^T u_i = 1$. We can rewrite this decomposition (using the basis defined by the eigenvectors) as:

$$\Sigma = \sum_i \lambda_i u_i u_i^T \tag{1.2}$$

The inverse of this matrix is then given by:

$$\Sigma^{-1} = \sum_i \frac{1}{\lambda_i} u_i u_i^T \tag{1.3}$$

This allows us to rewrite the Mahalanobis distance as follows:

$$\Delta \equiv D_M\left(x \mid \mu, \Sigma\right)^2 \tag{1.4}$$

$$= \sum_i \frac{1}{\lambda_i}\left(x-\mu\right)^T u_i u_i^T \left(x-\mu\right) \tag{1.5}$$

$$= \sum_i \frac{\left(u_i^T\left(x-\mu\right)\right)^2}{\lambda_i} \equiv \sum_i \frac{y_i^2}{\lambda_i} \tag{1.6}$$

where we defined $y_i = u_i^T\left(x-\mu\right)$. Notice that the density will be constant on the surfaces where Δ is constant. The shape described by 1.4 is an *ellipse* with radii equal to $\lambda_i^{1/2}$, centered around μ. This is really clear in the 2D case:

$$\Delta = \left(\frac{u_1^T\left(x-\mu\right)}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{u_2^T\left(x-\mu\right)}{\sqrt{\lambda_2}}\right)^2 \tag{1.7}$$

So in 2D, all of the *contour lines* (which are lines that have the same density along the PDF) will always be ellipses; in the multivariate case they will be ellipsoids (which is an ellipse in more dimensions, kind of). Figure 2 (right) shows this explicitly - the Gaussian is centered around the mean μ, the contour of $\Delta = 1$ is an ellipse with axes aligned and scaled by the eigenvectors and square root of the eigenvalues of the covariance matrix.

2 The Derivative Trick

The Gaussian distribution is, *by definition*, any distribution that is the exponent of a quadratic function, i.e. any distribution of the form:

$$p\left(x\right) \propto \exp\left[-x^T\Gamma x + b^T x + c\right] \tag{2.1}$$

is Gaussian (even though it doesn't seem like it at first). In this course, we will see distributions with a form similar to the above, but will want to find the actual parameters (μ and Σ) that define the Gaussian, instead of leaving it as it is written above.

If we know ahead of time that $p\left(x\right)$ is a Gaussian, we can leverage the following property of the Gaussian distribution:

$$\mu = \arg\max_x \mathcal{N}\left(x \mid \mu, \Sigma\right) \tag{2.2}$$

$$= \arg\min_x \left\{-\log\mathcal{N}\left(x \mid \mu, \Sigma\right)\right\} \tag{2.3}$$

$$= \arg\min_x \left\{\frac{1}{2}\Delta + \text{const}\right\} \tag{2.4}$$

Luckily, the Mahalanobis distance is a convex function, and only has one minima (it *is* a parabola, after all). That means that we can find the mean by finding the derivative of $-\log p\left(x\right)$ and equating to zero.

However, we can go even further:

$$\Delta = \frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)$$

$$\Leftrightarrow \frac{\partial\Delta}{\partial x} = \Sigma^{-1}\left(x-\mu\right) \tag{2.5}$$

$$\frac{\partial^2\Delta}{\partial x \partial x^T} = \Sigma^{-1} \tag{2.6}$$

So, if we know that $p\left(x\right)$ is a Gaussian, and we want to find μ and Σ, we can simply differentiate $-\log p\left(x\right)$ and try to manipulate the resulting expression until we get:

$$\frac{\partial}{\partial x}\left(-\log p\left(x\right)\right) = \Sigma^{-1}\left(x-\mu\right) \tag{2.7}$$

Example: Conditional distribution of a Gaussian

An important property of the multivariate Gaussian distribution is that if two sets of variables are jointly Gaussian, then the conditional distribution of one set on the other is also Gaussian.
Consider a Gaussian variable separated into 2 parts:

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \tag{2.8}$$

such that $x \sim \mathcal{N}(\mu, \Sigma)$. We can divide the mean and the covariance in a fitting manner:

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \tag{2.9}$$

Since the covariance is symmetrical, we know that $\Sigma_{ab} = \Sigma_{ba}^T$ and that Σ_{aa} and Σ_{bb} are symmetrical. Actually, it will be easier to use the precision matrix $\Lambda = \Sigma^{-1}$ and divide it up in the same manner:

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \tag{2.10}$$

Note that $\Lambda_{aa} \neq \Sigma_{aa}^{-1}$! Later we will find out how each part of Λ relates to each part of Σ.
Let's start by finding an expression for the conditional distribution $p(x_a|x_b)$. We can find this distribution by evaluating the distribution of $p(x_a, x_b)$ while fixing x_b to a certain value and re-normalizing (the conditional distribution is a legal distribution). We will start by rewriting the quadratic term and it's parts:

$$\begin{aligned} -\frac{1}{2}(x-\mu)^T \Lambda (x-\mu) = -\frac{1}{2}\Big[&(x_a-\mu_a)^T \Lambda_{aa}(x_a-\mu_a) + (x_a-\mu_a)^T \Lambda_{ab}(x_b-\mu_b) \\ &+ (x_b-\mu_b)^T \Lambda_{ba}(x_a-\mu_a) + (x_b-\mu_b)^T \Lambda_{bb}(x_b-\mu_b)\Big] \end{aligned} \tag{2.11}$$

This is still a quadratic expression w.r.t. x_a, so the conditional distribution $p(x_a|x_b)$ will also be Gaussian. Because the form of the Gaussian is not very flexible, as long as we find what the quadratic term is equal (the one in the exponent), the normalization will work itself out (since the conditional is also a distribution that must integrate up to 1).

We can now use the derivative trick! Defining:

$$\Delta = \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \tag{2.12}$$

Starting with the first derivative:

$$\frac{\partial \Delta}{\partial x_a} = \Lambda_{aa}(x_a-\mu_a) + \Lambda_{ab}(x_b-\mu_b) \tag{2.13}$$

The second derivative will give us:

$$\frac{\partial^2 \Delta}{\partial x_a \partial x_a^T} = \Lambda_{aa} \tag{2.14}$$

So we know that the covariance is equal to $\Sigma_{a|b} = \Lambda_{aa}^{-1}$. Using this new found knowledge, we can find the mean, if we can rewrite equation 2.13 as $\Lambda_{aa}\left(x_a - \mu_{a|b}\right)$ for some $\mu_{a|b}$. Let's try to do this. Recall that Λ_{aa} is invertible, so we can write:

$$\begin{aligned} \Lambda_{aa}(x_a-\mu_a) + \Lambda_{ab}(x_b-\mu_b) &= \Lambda_{aa}\left(x_a - \mu_a + \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\right) \\ &= \Lambda_{aa}\left[x_a - \left(\mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\right)\right] \\ &\triangleq \Lambda_{aa}\left(x_a - \mu_{a|b}\right) \end{aligned} \tag{2.15}$$

Which means that the conditional distribution is parameterized by:

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b) \tag{2.16}$$

$$\Sigma_{a|b} = \Lambda_{aa}^{-1} \tag{2.17}$$

Okay, now all that remains is to find what Λ_{aa} and Λ_{ab} are equal to in terms of Σ. To do this, we will use the following identity for partitioned matrices:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}^{-1} \tag{2.18}$$

where $M = \left(A - BD^{-1}C\right)^{-1}$ is called the *Schur component* of the matrix with respect to the sub-matrix D. Using our earlier partitioning, we get:

$$\Lambda_{aa} = \left(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1} \tag{2.19}$$

$$\Lambda_{ab} = -\Lambda_{aa}\Sigma_{ab}\Sigma_{bb}^{-1} \tag{2.20}$$

Finally, we have the expressions needed to describe the conditional distribution:

$$\begin{aligned} p\left(x_a|x_b\right) &= \mathcal{N}\left(x_a \,|\, \mu_{a|b}, \Sigma_{a|b}\right) \\ \mu_{a|b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}\left(x_b - \mu_b\right) \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \end{aligned} \tag{2.21}$$

Note that in this case, the conditional distribution is much easier to describe in terms of the precision matrix instead of the covariance matrix. When implementing the code for this, it may be simpler to save the precision matrix (as well as the covariance matrix) in memory to easily compute the conditional distribution.

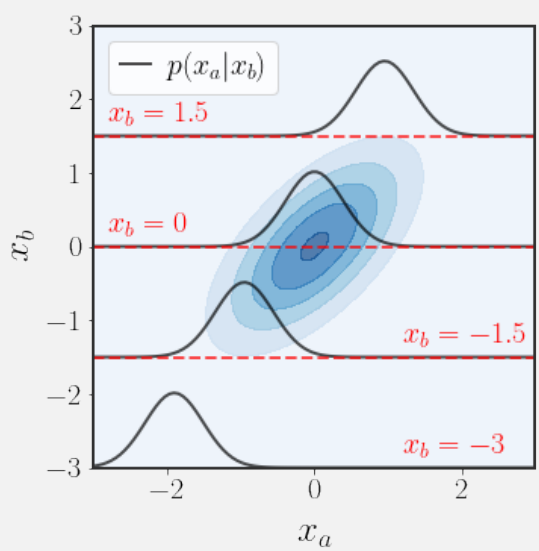

Figure 3: visualization of the conditional of a bivariate Gaussian; plots of $p\left(x_a|x_b\right)$ for various values of x_b. Notice how the variance doesn't change for different values of x_b, only the mean of the conditional.

3 Completing the Squares

While the derivative trick is very useful, we can't always use it, since we might lose information that we want to keep when differentiating. In such cases, we can use a different trick - completing the squares.

"Completing the squares" means that we want to turn a quadratic *function* into the quadratic *form* (plus some residuals). Suppose we have the following expression:

$$f\left(x\right) = x^T A x + 2x^T b + c \tag{3.1}$$

In this case, completing the squares means we would like to bring $f\left(x\right)$ to the form:

$$f\left(x\right) = \underbrace{\left(x + \boxed{?}\right)^T \boxed{\boxed{?}} \left(x + \boxed{?}\right)}_{\text{depends on } x} + \underbrace{g\left(A, b, c\right)}_{\text{const w.r.t } x} \tag{3.2}$$

where $\boxed{?}$ stands in for some vector and $\boxed{\boxed{?}}$ stands in for some matrix. For the case presented above, we can do so in the following manner (assuming A is invertible[2]):

$$
\begin{aligned}
f(x) &= x^T A x + 2x^T b + c && (3.3)\\
&= x^T A x + 2x^T A A^{-1} b + c && (3.4)\\
&= x^T A x + 2x^T A A^{-1} b - b^T A^{-1} A A^{-1} b + b^T A^{-1} A A^{-1} b + c && (3.5)\\
&= \left(x + A^{-1} b\right)^T A \left(x + A^{-1} b\right) - \underbrace{b^T A^{-1} b + c}_{\text{const w.r.t } x} && (3.6)
\end{aligned}
$$

Having written down the full quadratic form in 3.6, we can now understand which terms we lose when we use the derivative trick. When we differentiate $f(\cdot)$ with respect to x, we willingly drop all of the terms that are constant with respect to x - in this case, we would lose all information regarding $g(A, b, c)$. For example, suppose we want to find:

$$
p(y) \propto \int \exp\left[-\frac{1}{2}\left(x^T \Gamma x + 2x^T h(y)\right)\right] dx \tag{3.7}
$$

If we use the derivative trick to find the form of the Gaussian in the exponent, we would lose all information regarding y! This information is obviously important - we want to find $p(y)$, after all. Instead, plugging into the formula from 3.6, we have:

$$
\begin{aligned}
\int \exp\left[-\frac{1}{2}\left(x^T \Gamma x + 2x^T h(y)\right)\right] dx &= \int \exp\left[-\frac{1}{2}\left(x - \Gamma^{-1} h(y)\right)^T \Gamma \left(x - \Gamma^{-1} h(y)\right) + \frac{1}{2} h(y)^T \Gamma^{-1} h(y)\right] dx && (3.8)\\
&\propto e^{\frac{1}{2} h(y)^T \Gamma^{-1} h(y)} \int \mathcal{N}\left(x \,|\, \Gamma^{-1} h(y), \Gamma^{-1}\right) dx && (3.9)\\
&= e^{\frac{1}{2} h(y)^T \Gamma^{-1} h(y)} \propto p(y) && (3.10)
\end{aligned}
$$

Example: Marginal distribution of a Gaussian

Another important property of the Gaussian distribution is that its marginals are also Gaussian, which is what we will show in this example.
Again, we consider a Gaussian variable separated into 2 parts:

$$
\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right)
$$

and we will again define $\Lambda = \Sigma^{-1}$. We want to find $p(x_a)$.

Our battle plan is to first find all of the dependence on x_b, and to integrate it out. If we can do this without losing track of x_a, we will be able to find the marginal distribution and win. Notice that the Mahalanobis distance is quadratic in x_b (as usual), so the final form we will get to will be something like:

$$
p(x_a) = f(x_a) \int \mathcal{N}\left(x_b \,|\, \mu_{x_b}, \Sigma_{x_b}\right) dx_b = f(x_a)
$$

So, let's try to open up the Mahalanobis distance (the part in the exponent) and separate it into two groups: terms that contain x_b and those that don't.

[2] We can also do this when A is not invertible, in which case we will need to use the pseudo-inverse of A such that $AA^\dagger = I$

We begin by defining $y = x - \mu$, which in this case will allow us to write the Mahalanobis distance as:

$$\begin{aligned}
\Delta &= \frac{1}{2}\begin{pmatrix} y_a \\ y_b \end{pmatrix}^T \begin{pmatrix} \Lambda_a & B \\ B^T & \Lambda_b \end{pmatrix} \begin{pmatrix} y_a \\ y_b \end{pmatrix} \\
&= \frac{1}{2}\begin{pmatrix} y_a \\ y_b \end{pmatrix}^T \begin{pmatrix} \Lambda_a y_a + B y_b \\ B^T y_a + \Lambda_b y_b \end{pmatrix} \\
&= \frac{1}{2} y_a^T \Lambda_a y_a + \underbrace{\frac{1}{2}\left[2 y_b^T B y_a + y_b^T \Lambda_b y_b\right]}_{(*)}
\end{aligned} \tag{3.11}$$

Note that if we find the marginal of y_a, we effectively find the marginal of x_a, only we don't have to keep track of μ_a and μ_b! We now need to complete the squares in $(*)$ to find the complete dependence on y_b:

$$\begin{aligned}
2 y_b^T B y_a + y_b^T \Lambda_b y_b &= y_b^T \Lambda_b y_b + 2 y_b^T B y_a \\
&= y_b^T \Lambda_b y_b + 2 y_b^T \Lambda_b \Lambda_b^{-1} B y_a \\
&= y_b^T \Lambda_b y_b + 2 y_b^T \Lambda_b \Lambda_b^{-1} B y_a + y_a^T B^T \Lambda_b^{-1} \Lambda_b \Lambda_b^{-1} B y_a - y_a^T B^T \Lambda_b^{-1} \Lambda_b \Lambda_b^{-1} B y_a \\
&= \left(y_b + \Lambda_b^{-1} B y_a\right)^T \Lambda_b \left(y_b + \Lambda_b^{-1} B y_a\right) - y_a^T B^T \Lambda_b^{-1} B y_a
\end{aligned} \tag{3.12}$$

We add this back to the full Mahalanobis distance to get:

$$\Delta = \frac{1}{2} y_a^T \left(\Lambda_a - B^T \Lambda_b^{-1} B\right) y_a + \frac{1}{2}\left(y_b + \Lambda_b^{-1} B y_a\right)^T \Lambda_b \left(y_b + \Lambda_b^{-1} B y_a\right) \tag{3.13}$$

So our distribution is:

$$\begin{aligned}
p\left(y_a\right) &\propto \exp\left[-\frac{1}{2} y_a^T \left(\Lambda_a - B^T \Lambda_b^{-1} B\right) y_a\right] \int \exp\left[-\frac{1}{2}\left(y_b + \Lambda_b^{-1} B y_a\right)^T \Lambda_b \left(y_b + \Lambda_b^{-1} B y_a\right)\right] dy_b \\
&\propto \exp\left[-\frac{1}{2} y_a^T \left(\Lambda_a - B^T \Lambda_b^{-1} B\right) y_a\right]
\end{aligned} \tag{3.14}$$

which is definitely Gaussian! We were allowed to do the second move because the integral will be the normalization term of the Gaussian, which is a function of Λ_b- which is constant with respect to y_a (and so is eaten up by the $\propto$ sign).

Finally, we see that:

$$y_a \sim \mathcal{N}\left(0, \left(\Lambda_a - B^T \Lambda_b^{-1} B\right)^{-1}\right) \Rightarrow x_a \sim \mathcal{N}\left(\mu_a, \left(\Lambda_a - B^T \Lambda_b^{-1} B\right)^{-1}\right) \tag{3.15}$$

and all that remains is to find what the covariance is equal to in terms of Σ. To do this, we use the same identity we saw in the previous example:

$$\Sigma_{aa} = \left(\Lambda_{aa} - B^T \Lambda_b^{-1} B\right)^{-1} \tag{3.16}$$

So, the marginal is:

$$x_a \sim \mathcal{N}\left(\mu_a, \Sigma_{aa}\right) \tag{3.17}$$

which really makes you wonder why we did all of that hard work.

4 Extras

We saw the so called "derivative trick" and how completing the squares can also be of help, but it might not be obvious when to use each approach. First, remember that whenever we see a distribution of the form:

$$p\left(x, y\right) \propto \exp\left[-x^T \Gamma x + b\left(y\right)^T x + g\left(y\right)\right] \tag{4.1}$$

then $p(x)$ and $p(x|y)$ will be Gaussians[3], and we will usually want to find the "Gaussian form" we know and love. Once we figured that out, we can try to ask "how can we find the Gaussian form?", and the answer will usually be one of the following methods:

- If we don't care about $p(y)$ or $p(y|x)$ at all, i.e. we want to specifically find $p(x)$ or $p(x|y)$, then we can use the derivative trick
- If we need to know $p(y)$ or $p(y|x)$ explicitly as well as $p(x)$ or $p(x|y)$, then completing the squares is usually the easiest way
- When all else fails, but we know that what we are looking for is Gaussian, we can calculate the expectations $\mathbb{E}[\cdot]$ and covariance $\mathrm{cov}[\cdot]$ explicitly, since a Gaussian is completely defined by these two values

Once you fully understand why each method works, it will become quite clear when you should use each of them.

[3] This is a slightly more general statement than what we showed in the recitation, but you can verify the validity for yourself in these cases as well