More on Multivariate Gaussians

Chuong B. Do

November 21, 2008

Up to this point in class, you have seen multivariate Gaussians arise in a number of applications, such as the probabilistic interpretation of linear regression, Gaussian discriminant analysis, mixture of Gaussians clustering, and most recently, factor analysis. In these lecture notes, we attempt to demystify some of the fancier properties of multivariate Gaussians that were introduced in the recent factor analysis lecture. The goal of these notes is to give you some intuition into where these properties come from, so that you can use them with confidence on your homework (hint hint!) and beyond.

1 Definition

A vector-valued random variable $x \in \mathbf{R}^n$ is said to have a **multivariate normal (or Gaussian) distribution** with mean $\mu \in \mathbf{R}^n$ and covariance matrix $\Sigma \in \mathbf{S}^n_{++}$[1] if its probability density function is given by

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

We write this as $x \sim \mathcal{N}(\mu, \Sigma)$.

2 Gaussian facts

Multivariate Gaussians turn out to be extremely handy in practice due to the following facts:

- **Fact #1:** If you know the mean μ and covariance matrix Σ of a Gaussian random variable x, you can write down the probability density function for x directly.

[1] Recall from the section notes on linear algebra that $\mathbf{S}^n_{++}$ is the space of symmetric positive definite $n \times n$ matrices, defined as

$$\mathbf{S}^n_{++} = \left\{A \in \mathbf{R}^{n\times n} : A = A^T \text{ and } x^T A x > 0 \text{ for all } x \in \mathbf{R}^n \text{ such that } x \neq 0\right\}.$$

- **Fact #2:** The following Gaussian integrals have closed-form solutions:

$$\int_{x \in \mathbf{R}^n} p(x; \mu, \Sigma) dx = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x; \mu, \Sigma) dx_1 \dots dx_n = 1$$
$$\int_{x \in \mathbf{R}^n} x_i p(x; \mu, \sigma^2) dx = \mu_i$$
$$\int_{x \in \mathbf{R}^n} (x_i - \mu_i)(x_j - \mu_j) p(x; \mu, \sigma^2) dx = \Sigma_{ij}.$$

- **Fact #3:** Gaussians obey a number of *closure* properties:
  - The sum of independent Gaussian random variables is Gaussian.
  - The marginal of a joint Gaussian distribution is Gaussian.
  - The conditional of a joint Gaussian distribution is Gaussian.

At first glance, some of these facts, in particular facts #1 and #2, may seem either intuitively obvious or at least plausible. What is probably not so clear, however, is why these facts are so powerful. In this document, we'll provide some intuition for how these facts can be used when performing day-to-day manipulations dealing with multivariate Gaussian random variables.

3 Closure properties

In this section, we'll go through each of the closure properties described earlier, and we'll either prove the property using facts #1 and #2, or we'll at least give some type of intuition as to why the property is true.

The following is a quick roadmap of what we'll cover:

	sums	marginals	conditionals
why is it Gaussian?	no	yes	yes
resulting density function	yes	yes	yes

3.1 Sum of independent Gaussians is Gaussian

The formal statement of this rule is:

> Suppose that $y \sim \mathcal{N}(\mu, \Sigma)$ and $z \sim \mathcal{N}(\mu', \Sigma')$ are independent Gaussian distributed random variables, where $\mu, \mu' \in \mathbf{R}^n$ and $\Sigma, \Sigma' \in \mathbf{S}^n_{++}$. Then, their sum is also Gaussian:
>
> $$y + z \sim \mathcal{N}(\mu + \mu', \Sigma + \Sigma').$$

Before we prove anything, here are some observations:

1. The first thing to point out is that the importance of the independence assumption in the above rule. To see why this matters, suppose that $y \sim \mathcal{N}(\mu, \Sigma)$ for some mean vector μ and covariance matrix Σ, and suppose that $z = -y$. Clearly, z also has a Gaussian distribution (in fact, $z \sim \mathcal{N}(-\mu, \Sigma)$, but $y + z$ is identically zero!

2. The second thing to point out is a point of confusion for many students: if we add together two Gaussian densities ("bumps" in multidimensional space), wouldn't we get back some bimodal (i.e., "two-humped" density)? Here, the thing to realize is that the density of the random variable $y + z$ in this rule is NOT found by simply adding the densities of the individual random variables y and z. Rather, the density of $y + z$ will actually turn out to be a *convolution* of the densities for y and z.[2] To show that the convolution of two Gaussian densities gives a Gaussian density, however, is beyond the scope of this class.

Instead, let's just use the observation that the convolution does give some type of Gaussian density, along with Fact #1, to figure out what the density, $p(y + z|\mu, \Sigma)$ would be, if we were to actually compute the convolution. How can we do this? Recall that from Fact #1, a Gaussian distribution is fully specified by its mean vector and covariance matrix. If we can determine what these are, then we're done.

But this is easy! For the mean, we have

$$E[y_i + z_i] = E[y_i] + E[z_i] = \mu_i + \mu'_i$$

from linearity of expectations. Therefore, the mean of $y + z$ is simply $\mu + \mu'$. Also, the (i, j)th entry of the covariance matrix is given by

$$\begin{aligned}
&E[(y_i + z_i)(y_j + z_j)] - E[y_i + z_i]E[y_j + z_j] \\
&\quad = E[y_iy_j + z_iy_j + y_iz_j + z_iz_j] - (E[y_i] + E[z_i])(E[y_j] + E[z_j]) \\
&\quad = E[y_iy_j] + E[z_iy_j] + E[y_iz_j] + E[z_iz_j] - E[y_i]E[y_j] - E[z_i]E[y_j] - E[y_i]E[z_j] - E[z_i][z_j] \\
&\quad = (E[y_iy_j] - E[y_i]E[y_j]) + (E[z_iz_j] - E[z_i]E[z_j]) \\
&\qquad + (E[z_iy_j] - E[z_i]E[y_j]) + (E[y_iz_j] - E[y_i]E[z_j]).
\end{aligned}$$

Using the fact that y and z are independent, we have $E[z_iy_j] = E[z_i]E[y_j]$ and $E[y_iz_j] = E[y_i]E[z_j]$. Therefore, the last two terms drop out, and we are left with,

$$\begin{aligned}
&E[(y_i + z_i)(y_j + z_j)] - E[y_i + z_i]E[y_j + z_j] \\
&\quad = (E[y_iy_j] - E[y_i]E[y_j]) + (E[z_iz_j] - E[z_i]E[z_j]) \\
&\quad = \Sigma_{ij} + \Sigma'_{ij}.
\end{aligned}$$

[2] For example, if y and z were univariate Gaussians (i.e., $y \sim \mathcal{N}(\mu, \sigma^2)$, $z \sim \mathcal{N}(\mu', \sigma'^2)$), then the convolution of their probability densities is given by

$$\begin{aligned}
p(y + z; \mu, \mu', \sigma^2, \sigma'^2) &= \int_{-\infty}^{\infty} p(w; \mu, \sigma^2) p(y + z - w; \mu', \sigma'^2) dw \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(w - \mu)^2\right) \cdot \frac{1}{\sqrt{2\pi}\sigma'} \exp\left(-\frac{1}{2\sigma'^2}(y + z - w - \mu')^2\right) dw
\end{aligned}$$

From this, we can conclude that the covariance matrix of $y + z$ is simply $\Sigma + \Sigma'$.

At this point, take a step back and think about what we have just done. Using some simple properties of expectations and independence, we have computed the mean and covariance matrix of $y + z$. Because of Fact #1, we can thus write down the density for $y + z$ immediately, without the need to perform a convolution![3]

3.2 Marginal of a joint Gaussian is Gaussian

The formal statement of this rule is:

> Suppose that
>
> $$\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right),$$
>
> where $x_A \in \mathbf{R}^m$, $x_B \in \mathbf{R}^n$, and the dimensions of the mean vectors and covariance matrix subblocks are chosen to match x_A and x_B. Then, the marginal densities,
>
> $$\begin{aligned} p(x_A) &= \int_{x_B \in \mathbf{R}^n} p(x_A, x_B; \mu, \Sigma) dx_B \\ p(x_B) &= \int_{x_A \in \mathbf{R}^m} p(x_A, x_B; \mu, \Sigma) dx_A \end{aligned}$$
>
> are Gaussian:
>
> $$\begin{aligned} x_A &\sim \mathcal{N}(\mu_A, \Sigma_{AA}) \\ x_B &\sim \mathcal{N}(\mu_B, \Sigma_{BB}). \end{aligned}$$

To justify this rule, let's just focus on the marginal distribution with respect to the variables x_A.[4]

First, note that computing the mean and covariance matrix for a marginal distribution is easy: simply take the corresponding subblocks from the mean and covariance matrix of the joint density. To make sure this is absolutely clear, let's look at the covariance between $x_{A,i}$ and $x_{A,j}$ (the ith component of x_A and the jth component of x_A). Note that $x_{A,i}$ and $x_{A,j}$ are also the ith and jth components of

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix}$$

[3] Of course, we needed to know that $y + z$ had a Gaussian distribution in the first place.

[4] In general, for a random vector x which has a Gaussian distribution, we can always permute entries of x so long as we permute the entries of the mean vector and the rows/columns of the covariance matrix in the corresponding way. As a result, it suffices to look only at x_A, and the result for x_B follows immediately.

(since x_A appears at the top of this vector). To find their covariance, we need to simply look at the (i, j)th element of the covariance matrix,

$$\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$$

The (i, j)th element is found in the Σ_{AA} subblock, and in fact, is precisely $\Sigma_{AA,ij}$. Using this argument for all $i, j \in \{1, \ldots, m\}$, we see that the covariance matrix for x_A is simply Σ_{AA}. A similar argument can be used to find that the mean of x_A is simply μ_A. Thus, the above argument tells us that if we knew that the marginal distribution over x_A is Gaussian, then we could immediately write down a density function for x_A in terms of the appropriate submatrices of the mean and covariance matrices for the joint density!

The above argument, though simple, however, is somewhat unsatisfying: how can we actually be sure that x_A has a multivariate Gaussian distribution? The argument for this is slightly long-winded, so rather than saving up the punchline, here's our plan of attack up front:

1. Write the integral form of the marginal density explicitly.
2. Rewrite the integral by partitioning the inverse covariance matrix.
3. Use a "completion-of-squares" argument to evaluate the integral over x_B.
4. Argue that the resulting density is Gaussian.

Let's see each of these steps in action.

3.2.1 The marginal density in integral form

Suppose that we wanted to compute the density function of x_A directly. Then, we would need to compute the integral,

$$\begin{aligned} p(x_A) &= \int_{x_B \in \mathbf{R}^n} p(x_A, x_B; \mu, \Sigma) dx_B \\ &= \frac{1}{(2\pi)^{\frac{m+n}{2}} \begin{vmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{vmatrix}^{1/2}} \int_{x_B \in \mathbf{R}^n} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) dx_B. \end{aligned}$$

3.2.2 Partitioning the inverse covariance matrix

To make any sort of progress, we'll need to write the matrix product in the exponent in a slightly different form. In particular, let us define the matrix $V \in \mathbf{R}^{(m+n)\times(m+n)}$ as[5]

$$V = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \Sigma^{-1}.$$

[5]Sometimes, V is called the "precision" matrix.

It might be tempting to think that

$$V = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \text{“=”} \begin{bmatrix} \Sigma_{AA}^{-1} & \Sigma_{AB}^{-1} \\ \Sigma_{BA}^{-1} & \Sigma_{BB}^{-1} \end{bmatrix}$$

However, the rightmost equality does not hold! We'll return to this issue in a later step; for now, though, it suffices to define V as above without worrying what actual contents of each submatrix are.

Using this definition of V, the integral expands to

$$p(x_A) = \frac{1}{Z} \int_{x_B \in \mathbf{R}^n} \exp\Big(- \Big[\frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) + \frac{1}{2}(x_A - \mu_A)^T V_{AB}(x_B - \mu_B) + \frac{1}{2}(x_B - \mu_B)^T V_{BA}(x_A - \mu_A) + \frac{1}{2}(x_B - \mu_B)^T V_{BB}(x_B - \mu_B) \Big] \Big) dx_B,$$

where Z is some constant not depending on either x_A or x_B that we'll choose to ignore for the moment. If you haven't worked with partitioned matrices before, then the expansion above may seem a little magical to you. It is analogous to the idea that when defining a quadratic form based on some 2×2 matrix A, then

$$x^T A x = \sum_i \sum_j A_{ij} x_i x_j = x_1 A_{11} x_1 + x_1 A_{12} x_2 + x_2 A_{21} x_1 + x_2 A_{22} x_2.$$

Take some time to convince yourself that the matrix generalization above also holds.

3.2.3 Integrating out x_B

To evaluate the integral, we'll somehow want to integrate out x_B. In general, however, Gaussian integrals are hard to compute by hand. Is there anything we can do to save time? There are, in fact, a number of Gaussian integrals for which the answer is already known (see Fact #2). The basic idea in this section, then, will be to transform the integral we had in the last section into a form where we can apply one of the results from Fact #2 in order to perform the required integration easily.

The key to this is a mathematical trick known as "completion of squares." Consider the quadratic function $z^T Az + b^T z + c$ where A is a symmetric, nonsingular matrix. Then, one can verify directly that

$$\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2}\big(z + A^{-1}b\big)^T A \big(z + A^{-1}b\big) + c - \frac{1}{2} b^T A^{-1} b.$$

This is the multivariate generalization of the "completion of squares" argument used in single variable algebra:

$$\frac{1}{2} a z^2 + bz + c = \frac{1}{2} a \left(z + \frac{b}{a} \right)^2 + c - \frac{b^2}{2a}$$

To apply the completion of squares in our situation above, let

$$\begin{aligned}
z &= x_B - \mu_B \\
A &= V_{BB} \\
b &= V_{BA}(x_A - \mu_A) \\
c &= \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A).
\end{aligned}$$

Then, it follows that the integral can be rewritten as

$$\begin{aligned}
p(x_A) = \frac{1}{Z} \int_{x_B \in \mathbf{R}^n} \exp\Bigg(-\Bigg[&\frac{1}{2}\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big)^T V_{BB}\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big) \\
&+ \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) - \frac{1}{2}(x_A - \mu_A)^T V_{AB}V_{BB}^{-1}V_{BA}(x_A - \mu_A) \Bigg] \Bigg) dx_B
\end{aligned}$$

We can factor out the terms not including x_B to obtain,

$$\begin{aligned}
p(x_A) &= \exp\left(-\frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) + \frac{1}{2}(x_A - \mu_A)^T V_{AB}V_{BB}^{-1}V_{BA}(x_A - \mu_A) \right) \\
&\quad \cdot \frac{1}{Z} \int_{x_B \in \mathbf{R}^n} \exp\Bigg(-\frac{1}{2}\Bigg[\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big)^T V_{BB}\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big) \Bigg] \Bigg) dx_B
\end{aligned}$$

At this point, we can now apply Fact #2. In particular, we know that generically speaking, for a multivariate Gaussian distributed random variable x with mean μ and covariance matrix Σ, the density function normalizes, i.e.,

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \int_{\mathbf{R}^n} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right) = 1,$$

or equivalently,

$$\int_{\mathbf{R}^n} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right) = (2\pi)^{n/2}|\Sigma|^{1/2}.$$

We use this fact to get rid of the remaining integral in our expression for $p(x_A)$:

$$p(x_A) = \frac{1}{Z} \cdot (2\pi)^{n/2}|V_{BB}|^{1/2} \cdot \exp\left(-\frac{1}{2}(x_A - \mu_A)^T (V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})(x_A - \mu_A) \right).$$

3.2.4 Arguing that resulting density is Gaussian

At this point, we are almost done! Ignoring the normalization constant in front, we see that the density of x_A is the exponential of a quadratic form in x_A. We can quickly recognize that our density is none other than a Gaussian with mean vector μ_A and covariance matrix $(V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1}$. Although the form of the covariance matrix may seem a bit complex, we have already achieved what we set out to show in the first place—namely, that x_A has a marginal Gaussian distribution. Using the logic before, we can conclude that this covariance matrix must somehow reduce to Σ_{AA}.

But, in case you are curious, it's also possible to show that our derivation is consistent with this earlier justification. To do this, we use the following result for partitioned matrices:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M^{-1} & -M^{-1}BD^{-1} \\ -D^{-1}CM^{-1} & D^{-1} + D^{-1}CM^{-1}BD^{-1} \end{bmatrix}.$$

where $M = A - BD^{-1}C$. This formula can be thought of as the multivariable generalization of the explicit inverse for a 2×2 matrix,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Using the formula, it follows that

$$\begin{aligned}
\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} &= \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} (V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1} & -(V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1}V_{AB}V_{BB}^{-1} \\ -V_{BB}^{-1}V_{BA}(V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1} & (V_{BB} - V_{BA}V_{AA}^{-1}V_{AB})^{-1} \end{bmatrix}
\end{aligned}$$

We immediately see that $(V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1} = \Sigma_{AA}$, just as we expected!

3.3 Conditional of a joint Gaussian is Gaussian

The formal statement of this rule is:

> Suppose that
>
> $$\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right),$$
>
> where $x_A \in \mathbf{R}^m$, $x_B \in \mathbf{R}^n$, and the dimensions of the mean vectors and covariance matrix subblocks are chosen to match x_A and x_B. Then, the conditional densities
>
> $$\begin{aligned}
> p(x_A \mid x_B) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_A \in \mathbf{R}^m} p(x_A, x_B; \mu, \Sigma) dx_A} \\
> p(x_B \mid x_A) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B \in \mathbf{R}^n} p(x_A, x_B; \mu, \Sigma) dx_B}
> \end{aligned}$$
>
> are also Gaussian:
>
> $$\begin{aligned}
> x_A \mid x_B &\sim \mathcal{N}\big(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\big) \\
> x_B \mid x_A &\sim \mathcal{N}\big(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}\big).
> \end{aligned}$$

As before, we'll just examine the conditional distribution $x_B \mid x_A$, and the other result will hold by symmetry. Our plan of attack will be as follows:

1. Write the form of the conditional density explicitly.
2. Rewrite the expression by partitioning the inverse covariance matrix.
3. Use a "completion-of-squares" argument.
4. Argue that the resulting density is Gaussian.

Let's see each of these steps in action.

3.3.1 The conditional density written explicitly

Suppose that we wanted to compute the density function of x_B given x_A directly. Then, we would need to compute

$$\begin{aligned} p(x_B \mid x_A) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B \in \mathbf{R}^m} p(x_A, x_B; \mu, \Sigma) dx_A} \\ &= \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \end{aligned}$$

where Z' is a normalization constant that we used to absorb factors not depending on x_B. Note that this time, we don't even need to compute any integrals – the value of the integral does not depend on x_B, and hence the integral can be folded into the normalization constant Z'.

3.3.2 Partitioning the inverse covariance matrix

As before, we reparameterize our density using the matrix V, to obtain

$$\begin{aligned} p(x_B \mid x_A) &= \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \\ &= \frac{1}{Z'} \exp\Big(- \Big[\frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) + \frac{1}{2}(x_A - \mu_A)^T V_{AB}(x_B - \mu_B) \\ &\qquad + \frac{1}{2}(x_B - \mu_B)^T V_{BA}(x_A - \mu_A) + \frac{1}{2}(x_B - \mu_B)^T V_{BB}(x_B - \mu_B) \Big] \Big). \end{aligned}$$

3.3.3 Use a "completion of squares" argument

Recall that

$$\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2} \big(z + A^{-1} b\big)^T A \big(z + A^{-1} b\big) + c - \frac{1}{2} b^T A^{-1} b$$

provided A is a symmetric, nonsingular matrix. As before, to apply the completion of squares in our situation above, let

$$\begin{aligned}
z &= x_B - \mu_B \\
A &= V_{BB} \\
b &= V_{BA}(x_A - \mu_A) \\
c &= \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A).
\end{aligned}$$

Then, it follows that the expression for $p(x_B \mid x_A)$ can be rewritten as

$$\begin{aligned}
p(x_B \mid x_A) = \frac{1}{Z'} \exp\Bigg(-\Bigg[&\frac{1}{2}\big(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\big)^T V_{BB} \big(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\big) \\
&+ \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) - \frac{1}{2}(x_A - \mu_A)^T V_{AB} V_{BB}^{-1} V_{BA}(x_A - \mu_A) \Bigg] \Bigg)
\end{aligned}$$

Absorbing the portion of the exponent which does not depend on x_B into the normalization constant, we have

$$p(x_B \mid x_A) = \frac{1}{Z''} \exp\Bigg(-\frac{1}{2}\big(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\big)^T V_{BB} \big(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\big) \Bigg)$$

3.3.4 Arguing that resulting density is Gaussian

Looking at the last form, $p(x_B \mid x_A)$ has the form of a Gaussian density with mean $\mu_B - V_{BB}^{-1} V_{BA}(x_A - \mu_A)$ and covariance matrix V_{BB}^{-1}. As before, recall our matrix identity,

$$\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} = \begin{bmatrix} (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1} & -(V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1} V_{AB} V_{BB}^{-1} \\ -V_{BB}^{-1} V_{BA}(V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1} & (V_{BB} - V_{BA} V_{AA}^{-1} V_{AB})^{-1} \end{bmatrix}.$$

From this, it follows that

$$\mu_{B|A} = \mu_B - V_{BB}^{-1} V_{BA}(x_A - \mu_A) = \mu_B + \Sigma_{BA} \Sigma_{AA}^{-1}(x_A - \mu_A).$$

Conversely, we can also apply our matrix identity to obtain:

$$\begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} (\Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})^{-1} & -(\Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})^{-1} \Sigma_{AB} \Sigma_{BB}^{-1} \\ -\Sigma_{BB}^{-1} \Sigma_{BA}(\Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})^{-1} & (\Sigma_{BB} - \Sigma_{BA} \Sigma_{AA}^{-1} \Sigma_{AB})^{-1} \end{bmatrix},$$

from which it follows that

$$\Sigma_{B|A} = V_{BB}^{-1} = \Sigma_{BB} - \Sigma_{BA} \Sigma_{AA}^{-1} \Sigma_{AB}.$$

And, we're done!

4 Summary

In these notes, we used a few simple properties of multivariate Gaussians (plus a couple matrix algebra tricks) in order to argue that multivariate Gaussians satisfy a number of closure properties. In general, multivariate Gaussians are exceedingly useful representations of probability distributions because the closure properties ensure that most of the types of operations we would ever want to perform using a multivariate Gaussian can be done in closed form. Analytically, integrals involving multivariate Gaussians are often nice in practice since we can rely on known Gaussian integrals to avoid having to ever perform the integration ourselves.

5 Exercise

Test your understanding! Let $A \in \mathbf{R}^{n \times n}$ be a symmetric nonsingular square matrix, $b \in \mathbf{R}^n$, and c. Prove that

$$\int_{x \in \mathbf{R}^n} \exp\left(-\frac{1}{2}x^T A x - x^T b - c\right) dx = \frac{(2\pi)^{n/2}}{|A|^{1/2} \exp(c - b^T A^{-1} b)}.$$

References

For more information on multivariate Gaussians, see

Bishop, Christopher M. *Pattern Recognition and Machine Learning.* Springer, 2006.