# Handout Probabilistic Graphical Models: Gaussian Computations

Volkan Cevher, Matthias Seeger

**Abstract**

This is a loosely organized set of notes complementing the lecture on *Gaussian Distributions*. It is meant to help you doing Gaussian computations in an economic way. A word of warning: there is no easy way into this domain, there is just practice.

## 1 Definitions

The density of a multivariate Gaussian variable $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{x} \in \mathbb{R}^n$) is given as

$$P(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} e^{-(1/2)(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}. \tag{1}$$

Here, $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{x}]$ is the mean, $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{x}]$ is the covariance matrix of $P(\boldsymbol{x})$. $|\cdot|$ denotes the determinant of a matrix, the product of its eigenvalues.

A Gaussian is completely defined by mean and covariance. In the maximum entropy sense, it is "nothing else but mean and covariance", meaning that there is no additional structure. In *general* (not only for Gaussian variables), mean and covariance linearly transform in the following way. If $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{\mu}$, $\mathrm{Cov}[\boldsymbol{x}] = \boldsymbol{\Sigma}$, and $\boldsymbol{y} = \boldsymbol{E}\boldsymbol{x} + \boldsymbol{b}$, then

$$\mathrm{E}[\boldsymbol{y}] = \boldsymbol{E}\boldsymbol{\mu} + \boldsymbol{b}, \quad \mathrm{Cov}[\boldsymbol{y}] = \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{E}^T. \tag{2}$$

These properties hold simply because expectation is a linear transformation.

I will also frequently use the notation $A \doteq B$, which means that $A = CB$, where $C$ is some constant that does not matter right now. If $\doteq$ is used several times, the constants might be different.

## 2 Closure Properties. How to Determine a Gaussian Result

A family of distributions is *closed* under a set of operations on distributions or random variables if whenever you apply an operation to a family member, the outcome lies in the family as well. The Gaussian family is amazing in that sense.

- Gaussians are closed under linear (affine) transformations:

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \boldsymbol{y} = \boldsymbol{E}\boldsymbol{x} + \boldsymbol{b} \quad \Rightarrow \quad \boldsymbol{y} \sim N(\boldsymbol{E}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{E}^T)$$

- Gaussians are closed under marginalization:

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\ I \subset \{1, \ldots, n\} \quad \Rightarrow \quad \boldsymbol{x}_I = (x_i)_{i \in I} \text{ Gaussian}$$

  In other words: the *sum* rule retains Gaussianity.

- Gaussians are closed under conditioning:

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\ I \subset \{1, \ldots, n\},\ R = \{1, \ldots, n\} \setminus I \quad \Rightarrow \quad P(\boldsymbol{x}_I | \boldsymbol{x}_R) \text{ Gaussian}$$

  In other words: the *product* rule retains Gaussianity.

Beware the vocabulary. We say that a distribution is Gaussian, or that a random variable is Gaussian (the latter is sloppy for "is Gaussian distributed").

Here is a very important recipe for *determining a Gaussian result*, which often works even in complicated models or situations. You combine what you know about closure properties with what you know about transformations of mean and covariance:

1. Establish that the desired result has to be a Gaussian.
2. Determine its mean and covariance. Since a Gaussian is defined by these, you are done. In particular, there is *no need to track multiplicative constants*: from mean, covariance, and (1), you figure them out at the end.

For the first part, there is a number of ways you can try:

- Use closure properties, possibly in combination.
- Write down the functional form of the result, then massage it into the *Gaussian form*: $Ce^{-q(\boldsymbol{x})}$, $q(\boldsymbol{x})$ a quadratic function. If you manage this, it's a Gaussian.
- Use a sampling argument. If you can formulate the desired result as part of a joint sampling rule (say, a conditional distribution) where all input distributions are Gaussian and all transformation are affine, all intermediate distributions have to be Gaussian.

These points also help you with doing the second part (working out mean, covariance) in an economic fashion. A word of warning: there are books full of Gaussian formulae. You can learn them by heart and solve all these problems by long pages of algebra. But that's not fun, it's error-prone, and it's simply very embarassing in the long run. In the following, I will give examples for how to use these points, and then hope for really elegant solutions to exercise sheet problems.

## 2.1 Linear Transformations. Marginal Distributions

The first example is simple. Suppose $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{y} = \boldsymbol{E}\boldsymbol{x} + \boldsymbol{b}$. What is $P(\boldsymbol{y})$? First step: $P(\boldsymbol{y})$ is Gaussian, because Gaussians are closed under affine transformations, and we start with a Gaussian.

At this point, you might be a bit sceptical. What if $\boldsymbol{E}$ has a row that is zero? More generally, what if the rank of $\boldsymbol{E}$ is less than the number of rows? The covariance of $P(\boldsymbol{y})$ will then be

singular, and (1) cannot be true (the determinant is zero). This is an example of a *degenerate Gaussian*. Strictly speaking, most closure properties are true only if we extend the Gaussian family by these members. Geometrically, a degenerate Gaussian is just a Gaussian that lives in a space that is too high-dimensional for it to have mass there. It is like a flat pancake in this space. Although these pancake Gaussians are important to understand certain models later on, we can avoid them by always starting with *proper Gaussians* (positive definite covariance), and requiring that $\boldsymbol{E}$ in linear transforms has full rank (equal to number of rows).

Second step: use (2) to deduce mean and covariance of $P(\boldsymbol{y})$. Note that there is no need to compute the normalization constant of $P(\boldsymbol{y})$, because it is determined by mean and covariance.

What about the marginal distribution of $\boldsymbol{x}_I$, $I \subset \{1, \ldots, n\}$? That is a special case, since $\boldsymbol{x}_I = \boldsymbol{I}_{I,\cdot}\boldsymbol{x}$. A word on notation (more about this in Lecture 4). $\boldsymbol{x}_I = (x_i)_{i \in I}$, $I$ is an ordered index set. $\boldsymbol{A}_{I,J} = (a_{ij})_{i \in I, j \in J}$, $I$ for rows, $J$ for columns. I also use the shorthand $\boldsymbol{A}_I := \boldsymbol{A}_{I,I}$. *Selection matrices* are obtained by applying indices to the identity matrix. Make yourself clear that $\boldsymbol{x}_I = \boldsymbol{I}_{I,\cdot}\boldsymbol{x}$, or that $\boldsymbol{I}_{\cdot,I}\boldsymbol{x}_I$ is the vector of size $n$ which is equal to $\boldsymbol{x}$ at $I$, but zero elsewhere. Since marginalization is a special case of a linear transformation (for continuous variables), then $P(\boldsymbol{x}_I) = N(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$.

# 3 Conditional Distribution. Natural Parameterization

In this section, we will see that conditional Gaussians (product rule) are a bit harder to determine than marginal Gaussians (sum rule). In a sense, this is an artefact of the parameterization we use (mean, covariance): in other *natural* parameters, conditioning is equally simple. In practice, we have one or the other, so we need to know how to convert between them painlessly. In this section, I will demonstrate the general ideas of the previous section, thereby obtaining some important concepts (Schur complement) and identities (partitioned inverse equations, determinant identity, Woodbury formula).

## 3.1 Towards Natural Parameterization

Given that $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $I \subset \{1, \ldots, n\}$, let $R = \{1, \ldots, n\} \setminus I$. What is $P(\boldsymbol{x}_I | \boldsymbol{x}_R)$? Suppose we did not know that $P(\boldsymbol{x}_I | \boldsymbol{x}_R)$ is Gaussian. But we know that $P(\boldsymbol{x}) = P(\boldsymbol{x}_I, \boldsymbol{x}_R) = P(\boldsymbol{x}_I | \boldsymbol{x}_R) P(\boldsymbol{x}_R)$, so if we write down $P(\boldsymbol{x})$, the part depending on $\boldsymbol{x}_I$ will be the conditional distribution. From (1), we see that we have to partition the inverse: $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$. Then,

$$P(\boldsymbol{x}_I | \boldsymbol{x}_R) P(\boldsymbol{x}_R) \doteq e^{-(1/2)((\boldsymbol{x}_I - \boldsymbol{\mu}_I)^T \boldsymbol{A}_I (\boldsymbol{x}_I - \boldsymbol{\mu}_I)^T + 2(\boldsymbol{x}_R - \boldsymbol{\mu}_R)^T \boldsymbol{A}_{R,I} (\boldsymbol{x}_I - \boldsymbol{\mu}_I))}.$$

Note that I just write the quadratic and linear part w.r.t. $\boldsymbol{x}_I$. To determine what $P(\boldsymbol{x}_I | \boldsymbol{x}_R)$ is, I don't need to know the constant part: it'll be part of $P(\boldsymbol{x}_R)$. I *directly see* now that $P(\boldsymbol{x}_I | \boldsymbol{x}_R)$ must be Gaussian: it has the form $e^{-q(\boldsymbol{x}_I)}$, where $q$ is a quadratic function. Moreover, I can read off its mean and covariance by matching the expression against (1). First, $\mathrm{Cov}[\boldsymbol{x}_I | \boldsymbol{x}_R] = \boldsymbol{A}_I^{-1}$. Second,

$$(\boldsymbol{x}_R - \boldsymbol{\mu}_R)^T \boldsymbol{A}_{R,I} (\boldsymbol{x}_I - \boldsymbol{\mu}_I) = ((\boldsymbol{x}_R - \boldsymbol{\mu}_R)^T \boldsymbol{A}_{R,I} \boldsymbol{A}_I^{-1}) \boldsymbol{A}_I (\boldsymbol{x}_I - \boldsymbol{\mu}_I),$$

so that

$$\mathrm{E}[\boldsymbol{x}_I|\boldsymbol{x}_R] = \boldsymbol{\mu}_I + \boldsymbol{A}_I^{-1}\boldsymbol{A}_{I,R}(\boldsymbol{x}_R - \boldsymbol{\mu}_R).$$

It is a good exercise to multiply out the quadratic for $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$ from (1) and match terms against the expression for $P(\boldsymbol{x}_I|\boldsymbol{x}_R)P(\boldsymbol{x}_R)$ above, to retrieve these identities.

This means that while marginalization (sum rule) is simple when you know mean and covariance matrix, things are harder for the conditional distribution. However, they are *simple* again in a different parameterization. The one you know, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, is called *mean parameterization* (or *moment parameterization*. The novel one is obtained from this one as:

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}, \quad \boldsymbol{r} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

This is called *natural parameterization* (or *canonical parameterization*). The ultimate reason why Gaussian identities look somewhat messy, and unfortunately also why Gaussian information propagation is hard to do in a numerically stable way, is that we have to constantly convert between these parameterizations: inference involves both the sum and the product rule, and either is elementary in one of the parameterizations.

For a number of important reasons, among them numerical stability and processing time, it is *not* a good idea to constantly convert between parameterizations in practice. It's a useful way to think about the problem, but in practice you want to stay in one of the parameterizations and use direct identities. In the following, we will do that for the conditional distribution, given that we have mean parameters only. This way, you'll see some examples how these identities are obtained, illustrating ideas mentioned above already. We'll also derive some important general matrix identities this way, and see some "derivation saving" tricks in action.

### 3.2 Conditional Distribution by Sampling Argument

We derived mean and covariance of $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$ above, but the result is in terms of natural parameters. What if I only have mean parameters? I could convert them, but that involves a matrix inversion. We'll see in Lecture 4 how *bad* matrix inversion is in practice. It's numerically unstable and expensive. You want to avoid it whenever possible. This is not something they tell you in your linear algebra lectures, but it is *Numerical Mathematics 101*.

To get to $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$ directly, we can use a sampling argument. Let's first get rid of means by transforming $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu}$, adding it back in later. We know that $P(\boldsymbol{y}) = P(\boldsymbol{y}_I|\boldsymbol{y}_R)P(\boldsymbol{y}_R)$, which tells us how to sample $\boldsymbol{y}$:

1. Draw $\boldsymbol{y}_R \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_R)$
2. Draw $\boldsymbol{y}_I \sim P(\boldsymbol{y}_I|\boldsymbol{y}_R) = N(?, ?)$

And we know the outcome, namely $\boldsymbol{y} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. We use an *ansatz*, which means that we guess a form for the solution. The ansatz is that $\boldsymbol{y}_I = \boldsymbol{u} + \boldsymbol{B}\boldsymbol{y}_R$, where $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{C})$ is independent of $\boldsymbol{y}_R$. But this is just a fancy way of saying that $\boldsymbol{y}$ is a linear transform of $\boldsymbol{y}_R$ and $\boldsymbol{u}$:

$$\begin{bmatrix} \boldsymbol{y}_I \\ \boldsymbol{y}_R \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y}_R \end{bmatrix}. \tag{3}$$

We are on known ground now:

$$\mathrm{Cov}[\boldsymbol{y}] = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_R \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{B}^T & \boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} + \boldsymbol{B}\boldsymbol{\Sigma}_R\boldsymbol{B}^T & \boldsymbol{B}\boldsymbol{\Sigma}_R \\ \boldsymbol{B}^T\boldsymbol{\Sigma}_R & \boldsymbol{\Sigma}_R \end{bmatrix} \stackrel{!}{=} \begin{bmatrix} \boldsymbol{\Sigma}_I & \boldsymbol{\Sigma}_{I,R} \\ \boldsymbol{\Sigma}_{R,I} & \boldsymbol{\Sigma}_R \end{bmatrix} \tag{4}$$

Note that the matrix in the middle is block-diagonal, because $\boldsymbol{u}$ and $\boldsymbol{y}_R$ are independent. Now, just match terms to deduce $\boldsymbol{B}$ and $\boldsymbol{C}$. First, $\boldsymbol{B} = \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R^{-1}$. Second,

$$\boldsymbol{C} = \boldsymbol{\Sigma}_I - \boldsymbol{B}\boldsymbol{\Sigma}_R\boldsymbol{B}^T = \boldsymbol{\Sigma}_I - \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R^{-1}\boldsymbol{\Sigma}_{R,I} =: \boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R. \tag{5}$$

The form of $\boldsymbol{C}$ in terms of blocks of $\boldsymbol{\Sigma}$ is the *Schur complement*, which we revisit below.

Now what is $P(\boldsymbol{x}_I|\boldsymbol{y}_R)$? Just a word on why the ansatz and the sampling argument worked out. The ansatz is a rule for sampling $\boldsymbol{y}$. Since we start with Gaussians and do linear transforms, $\boldsymbol{y}$ must be Gaussian. Moreover, we managed to find $\boldsymbol{B}$ and $\boldsymbol{C}$ so that the "target parameters" $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are obtained. By the ansatz, $P(\boldsymbol{x}_I|\boldsymbol{y}_R)$ is Gaussian, and it is just one more step to its mean parameters. First, $\mathrm{E}[\boldsymbol{y}_I|\boldsymbol{y}_R] = \mathrm{E}[\boldsymbol{u} + \boldsymbol{B}\boldsymbol{y}_R|\boldsymbol{y}_R] = \boldsymbol{B}\boldsymbol{y}_R = \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R^{-1}(\boldsymbol{x}_R - \boldsymbol{\mu}_R)$. Second,

$$\mathrm{Cov}[\boldsymbol{x}_I|\boldsymbol{x}_R] = \mathrm{Cov}[\boldsymbol{u}] = \boldsymbol{C} = \boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R,$$

the Schur complement.

### 3.3 Schur Complement. Partitioned Inverse Equations

Recall that if $P(\boldsymbol{x}_I, \boldsymbol{x}_R) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the Schur complement (5) is the covariance matrix of the conditional distribution $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$. It exists and is invertible whenever $\boldsymbol{\Sigma}$ is invertible. The notation $\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R$ suggests something like "$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R)\boldsymbol{\Sigma}_R$", which does not hold of course. It does however hold for the normalization constants:

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R||\boldsymbol{\Sigma}_R|. \tag{6}$$

Namely, recall the decomposition (4) of $\boldsymbol{\Sigma}$, and note that

$$\left|\begin{bmatrix} \boldsymbol{I} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}\right| = 1,$$

since the determinant of a triangular matrix is the product of the diagonal entries. (6) follows, since the right hand side is the determinant of the block-diagonal matrix in the middle in (4). Another way to check (6) is to recall that $P(\boldsymbol{x}_I, \boldsymbol{x}_R) = P(\boldsymbol{x}_I|\boldsymbol{x}_R)P(\boldsymbol{x}_R)$, then to match normalization constants.

Note that $\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R$ is defined over arbitrary square invertible matrices, they do not have to be symmetric. (6) holds in general as well. For the special case

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{I} \end{bmatrix},$$

where the $\boldsymbol{I}$ can be of different size, we have that

$$|\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R| = |\boldsymbol{I} + \boldsymbol{U}\boldsymbol{V}| = |\boldsymbol{I} + \boldsymbol{V}\boldsymbol{U}| = |\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_I|,$$

a very useful determinant identity.

At this point, we have determined the conditional Gaussian in two different ways. We could match the results from Section 3.1 and Section 3.2 to obtain the partitioned inverse equations. But it is simpler to extend arguments from Section 3.2. Let's compute the inverse of the first equation in (4). Note that the matrices left and right have simple inverses:

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

Therefore,

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{B}^T & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} (\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_R^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \stackrel{!}{=} \begin{bmatrix} \boldsymbol{A}_I & \boldsymbol{A}_{I,R} \\ \boldsymbol{A}_{R,I} & \boldsymbol{A}_R \end{bmatrix},$$

where $\boldsymbol{B} = \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R^{-1}$. Multiplying these out (which you should do as exercise) results in the *partitioned inverse equations*. They just do what the name suggests. For a matrix $\boldsymbol{\Sigma}$, they are explicit formulae of blocks of the inverse in terms of corresponding blocks of $\boldsymbol{\Sigma}$. Once more, they hold for any nonsingular $\boldsymbol{\Sigma}$.

From the partitioned inverse equations, it is a simple step to the *Woodbury formula*. We simply apply the former with $I$ and $R$ interchanged, then equate terms. The significance of the Woodbury formula is that it allows you update the inverse of a matrix to which a low rank matrix is added. I strongly suggest that you derive the formula from the partitioned inverse equations as an exercise.

### 3.4 Another Sampling Argument

Here's another example for a sampling argument. We know that

$$N(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)N(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})C, \tag{7}$$

but what is $C$? Recall that in most cases, we don't have to know: all that matters is $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$. But sometimes we have to know, examples will come up in later lectures. I invite you to do this very messy computation by hand, then to appreciate the following simple argument.

We can see (7) as a statement about three variables: $\boldsymbol{x}$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$. Formally, it does not matter whether we write $N(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ or $N(\boldsymbol{\mu}_1|\boldsymbol{x}, \boldsymbol{\Sigma}_1)$. Suppose we take the integral over $\boldsymbol{x}$ on both sides. On the right side, we get $C$, since $\int N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x} = 1$. On the left side, flipping arguments in the first term, we have

$$\int N(\boldsymbol{\mu}_1|\boldsymbol{x}, \boldsymbol{\Sigma}_1)N(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\, d\boldsymbol{x}.$$

But this is a sampling recipe for the distribution of $\boldsymbol{\mu}_1$ given $\boldsymbol{\mu}_2$:

1. Sample $\boldsymbol{x} \sim N(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.
2. Sample $\boldsymbol{\mu}_1 \sim N(\boldsymbol{\mu}_1|\boldsymbol{x}, \boldsymbol{\Sigma}_1)$. Drop $\boldsymbol{x}$.

The combined sampling rule is $\boldsymbol{\mu}_1 = \boldsymbol{x} + \boldsymbol{v}$, $\boldsymbol{v} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$ independent of $\boldsymbol{x}$. But this means that $\boldsymbol{\mu}_1|\boldsymbol{\mu}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$, so that $C = N(\boldsymbol{\mu}_1|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.

# 4 Linear-Gaussian Model

Recall the linear-Gaussian model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{u} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Psi}).$$

This is a fundamental latent variable model: $\boldsymbol{u}$ is latent, $\boldsymbol{y}$ is observed. For modelling data, what matters is the marginal distribution of $\boldsymbol{y}$, whose structure is determined by the dimensionality of $\boldsymbol{u}$, its prior, the mapping $\boldsymbol{X}$, and the noise covariance $\boldsymbol{\Psi}$. We can use the *tower formulae* to compute the marginal distribution. Here is the proof for the covariance. Let $\boldsymbol{v} = \mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]$. Then,

$$\begin{aligned}\mathrm{E}[\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{u}]] &= \mathrm{E}\left[\mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T|\boldsymbol{u}] - \boldsymbol{v}\boldsymbol{v}^T\right] \stackrel{*}{=} \mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] = \mathrm{Cov}[\boldsymbol{y}] + \mathrm{E}[\boldsymbol{y}]\mathrm{E}[\boldsymbol{y}]^T - \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] \\ &\stackrel{*}{=} \mathrm{Cov}[\boldsymbol{y}] + \mathrm{E}[\boldsymbol{v}]\mathrm{E}[\boldsymbol{v}]^T - \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] = \mathrm{Cov}[\boldsymbol{y}] - \mathrm{Cov}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]].\end{aligned}$$

Here, we used the tower formulae for expectation at each point "*" (make sure you understand each of these applications). For the linear-Gaussian model, $\mathrm{E}[\boldsymbol{y}] = \mathrm{E}[\boldsymbol{X}\boldsymbol{u}] = \boldsymbol{X}\boldsymbol{\mu}_0$, while

$$\mathrm{Cov}[\boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{X}\boldsymbol{u}] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{u}]] = \mathrm{Cov}[\boldsymbol{X}\boldsymbol{u}] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{\varepsilon}]] = \boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T + \boldsymbol{\Psi}.$$

Another way is to obtain the joint distribution by using

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{X} \end{bmatrix} \boldsymbol{u} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\varepsilon} \end{bmatrix}$$

together with what we know about linear transforms of Gaussians. Note that $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are independent. For example, $\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}] = \boldsymbol{\Sigma}_0\boldsymbol{X}^T$.

Let us compute the posterior $P(\boldsymbol{u}|\boldsymbol{y})$ for this model, along the two different ways we derived above. We already know it must be Gaussian. Moreover,

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}] &= \mathrm{Cov}[(\boldsymbol{u}, \boldsymbol{y})]/\mathrm{Cov}[\boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{u}] - \mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}]\mathrm{Cov}[\boldsymbol{y}]^{-1}\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}]^T \\ &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T + \boldsymbol{\Psi})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}_0.\end{aligned}$$

Note that $\mathrm{Cov}[(\boldsymbol{u}, \boldsymbol{y})]$ is the joint covariance, while $\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}]$ is the cross-covariance (a block within the former). And

$$\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}] = \mathrm{E}[\boldsymbol{u}] + \mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}]\mathrm{Cov}[\boldsymbol{y}]^{-1}(\boldsymbol{y} - \mathrm{E}[\boldsymbol{y}]) = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_0\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T + \boldsymbol{\Psi})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_0).$$

We can also match terms in the joint distribution $P(\boldsymbol{y}, \boldsymbol{u}) = P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u})$. As shown in the lecture, this leads to

$$\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}, \quad \mathrm{E}[\boldsymbol{u}|\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0).$$

These are different expressions for the same thing. In general, the former set is more memorable and often easier to compute. The latter is more direct to obtain (you just have to match terms), but contains more inverses. Still, in some special cases, it might be the preferred way for this computation.

An important take-home message for the linear-Gaussian model is as follows. Suppose that $\boldsymbol{u} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$. Then, you can always compute posterior quantities by doing expensive (superlinear) computations, such as inverses, in the *smaller* number only: $\min\{m, n\}$. The main vehicle to formaly get from one set of expression to the other is the Woodbury formula.