ME 7247: Advanced Control Systems **Fall 2022–23**

Lecture 09: Marginal and Conditional Distributions

Friday October 9, 2022

Lecturer: Laurent Lessard *Scribe: Rushikesh Sankhye*

The previous lecture covered probability distributions, random vectors, Gaussian distributions and their transformations, expectations and covariance. This lecture covers marginal distributions, conditional distributions and joint distributions in the context of transformations on vectors in control and estimation problems.

1 Introduction

1.1 Recap

Consider a vector x that is normally distributed with mean μ and a covariance Σ:

$$x \sim \mathcal{N}(\mu, \Sigma)$$

An affine transformation of x will also be normally distributed, given by:

$$Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathsf{T})$$

1.2 Joint Distribution

If we have a larger vector that we partition into sub-vectors x and y, the distribution of $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ is called the *joint distribution*. Typically x and y will serve different role. For example, y may be a measurement, and x may be the variable we're trying to estimate. If the joint distribution is normal, and the density function is $f(x, y)$, we write:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_x \end{bmatrix} \right)$$

1.3 Marginal Distribution

Given a joint distribution (x, y), if we consider the distribution of x (just ignore y), this is called the *marginal distribution* of x. Similarly, the distribution of y is called the marginal distribution of y. We write these density functions as $f_x(x)$ and $f_y(y)$, respectively. They can be found by partially integrating the joint distribution:

$$f_x(x) = \int f(x, y)\,\mathrm{d}y$$
$$f_y(y) = \int f(x, y)\,\mathrm{d}x$$

1.4 Conditional Distribution

We can also consider the case where y is known, and we want to know what the distribution of x is *given* that y is what we measured it to be. We say this is the distribution of x *conditioned on* y. The conditional pdf is $f_{x|y}(x,y)$. Although it depends on y, it is a distribution in just x, so

$$\int f_{x|y}(x,y)\,\mathrm{d}x = 1 \qquad \text{for all } y. \tag{1}$$

The conditional distribution is a scaled version of the joint distribution. In order for the conditional pdf to integrate to 1, the scale factor must be the marginal with respect to y. That is,

$$f_{x|y}(x,y) = \frac{f(x,y)}{f_y(y)} \tag{2}$$

This is known as *Bayes' rule.*

2 Completing the square

The high-school approach for minimizing a quadratic function is called *completing the square.* We will derive a matrix version of this result. Let's start with a review of the standard scalar version. Say we have a quadratic of the form $ax^2 + 2bxy + dy^2 = 0$. The goal is to find the value of x that minimizes the expression. We do this by factoring out a (assuming $a \neq 0$), then manipulating the expression to make a square appear. Here are the steps:

$$\begin{aligned} ax^2 + 2bxy + dy^2 &= a\left(x^2 + \frac{2by}{a}x\right) + dy^2 \\ &= a\left(x + \frac{by}{a}\right)^2 - \frac{b^2y^2}{a} + dy^2 \\ &= a\left(x + \frac{by}{a}\right)^2 + \left(d - \frac{b^2}{a}\right)y^2 \end{aligned}$$

Only the first term depends on x. Depending on the sign of a, either it is positive (in which case the minimum is zero), or it is negative (in which case the minimum is unbounded).

$$x_{\text{opt}} = \begin{cases} \frac{-by}{a} & \text{if } a > 0 \\ \text{no solution} & \text{if } a < 0 \end{cases}$$

Note that the x that achieves the minimum value is a linear function of y. Meanwhile, the minimum value is a quadratic function of y given by $(d - \frac{b^2}{a})y^2$.

2.1 Completing squares: Matrix version

Given vectors x and y, the matrix version of completing squares can be written as

$$\underset{x}{\text{minimize}} \quad \begin{bmatrix} x \\ y \end{bmatrix}^\mathsf{T} \begin{bmatrix} A & B \\ B^\mathsf{T} & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x^\mathsf{T}Ax + 2x^\mathsf{T}By + y^\mathsf{T}Dy$$

Assuming A is invertible, the right hand side can be re-written as

$$\underset{x}{\text{minimize}} \quad \left(x + A^{-1}By\right)^{\mathsf{T}} A \left(x + A^{-1}By\right) + y^{\mathsf{T}} \left(D - B^{\mathsf{T}} A^{-1} B\right) y$$

To minimize this expression, substitute $x = -A^{-1}By$. The minimum value is $y^{\mathsf{T}}(D - B^{\mathsf{T}}A^{-1}B)y$. The optimal x is a *linear function* of y, and the optimal value is a *quadratic function* of y.

3 Matrix inversion lemma

Given a block 2×2 matrix where the $(1,1)$ and $(2,2)$ blocks are square, we can write down the following factorizations, depending on whether A is invertible, or D is invertible, respectively:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} \qquad \text{(block-LDU factorization)} \qquad (3a)$$

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix} \qquad \text{(block-UDL factorization)} \qquad (3b)$$

Now we can use the useful fact that block-upper or block-lower triangular matrices with identities in the diagonal blocks can be easily inverted:

$$\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix}$$

Inverting both sides of Eqs. (3a) and (3b), and applying the above formula, we obtain:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & \left(D - CA^{-1}B\right)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \qquad (4a)$$

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} \begin{bmatrix} \left(A - BD^{-1}C\right)^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \qquad (4b)$$

Expressions Eqs. (4a) and (4b) are two equivalent ways to compute the inverse of a block matrix. If we take the $(1,1)$ blocks of both sides, we obtain the well-known *Matrix Inversion Lemma* (MIL). It is also called the *Woodbury identity*, the *Sherman–Morrison formula*, or the *Sherman–Morrison–Woodbury identity.*

$$\left(A - BD^{-1}C\right)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1} CA^{-1} \qquad (5)$$

The MIL has many uses in computational linear algebra. The main use is when we have a matrix A whose inverse we have already calculated, and we would like to calculate the inverse of $A + uv^{\mathsf{T}}$ (a rank-one update to the matrix A). Without any help, we would have to calculate $(A + uv^{\mathsf{T}})^{-1}$ from scratch. It turns out we can leverage the fact that we already know A^{-1}. Applying MIL with $B = u$, $D = -1$, $C = v^{\mathsf{T}}$, we obtain:

$$\left(A + uv^{\mathsf{T}}\right)^{-1} = A^{-1} - A^{-1}u\left(1 + v^{\mathsf{T}}A^{-1}u\right)^{-1} v^{\mathsf{T}}A^{-1}$$

Now use the fact that $v^{\mathsf{T}} A^{-1} u$ is a *scalar* (1×1 matrix), so we can take its inverse by simple division. The result is:

$$\left(A + uv^{\mathsf{T}}\right)^{-1} = A^{-1} - \frac{1}{1 + v^{\mathsf{T}} A^{-1} u} \left(A^{-1} u\right) \left(v^{\mathsf{T}} A^{-1}\right)$$

In words, this says that the inverse of a rank-one update of A can be computed as a rank-one update to the inverse of A. This is called the *rank-one update formula.* We will later see that this formula is useful in recursive estimation.

4 Marginal Distribution

Let's assume that x and y have a joint Gaussian distribution.

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}\right) \tag{6}$$

Remember that covariance matrices are always positive definite. The marginal distribution is

$$f_x(x) = \int f(x,y)\,\mathrm{d}y$$

The expression for the joint distribution can be expanded as

$$f(x,y) = (\text{const}) \cdot \exp\left(-\frac{1}{2} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}^{-1} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}\right)$$

To lighten the notation, let $\tilde{x} := x - \mu_x$ and $\tilde{y} := y - \mu_y$.

$$f(x,y) = (\text{const}) \cdot \exp\left(-\frac{1}{2} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}\right)$$

Using the block-LDU factorization (4a), we can rewrite the joint distribution as:

$$\begin{aligned} f(x,y) &= (\text{const}) \cdot \exp\left(-\frac{1}{2} \begin{bmatrix} \tilde{x} \\ \tilde{y} - \Sigma_{yx}\Sigma_x^{-1}\tilde{x} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \Sigma_x^{-1} & 0 \\ 0 & (\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})^{-1} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} - \Sigma_{yx}\Sigma_x^{-1}\tilde{x} \end{bmatrix}\right) \\ &= (\text{const}) \cdot \exp\left(-\frac{1}{2}\tilde{x}^{\mathsf{T}}\Sigma_x^{-1}\tilde{x}\right) \cdot \exp\left(-\frac{1}{2}(\tilde{y} - \Sigma_{yx}\Sigma_x^{-1}\tilde{x})^{\mathsf{T}}\,(\Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})^{-1}(\tilde{y} - \Sigma_{yx}\Sigma_x^{-1}\tilde{x})\right) \end{aligned}$$

Now we evaluate $f_x(x) = \int f(x,y)\,\mathrm{d}y$. The first term does not contain y, so we can factor it out of the integral. The second term integrates to a constant (independent of x) because x only serves to shift the mean of the Gaussian pdf. We are integrating over the whole space of y's, so the shift does not affect the value of the integral. We conclude that

$$f_x(x) = (\text{const}) \cdot \exp\left(-\frac{1}{2}\tilde{x}^{\mathsf{T}}\Sigma_x^{-1}\tilde{x}\right)$$

Consequently, we have $x \sim \mathcal{N}(\mu_x, \Sigma_x)$.

4.1 Alternate Proof

Remember, if $x \sim \mathcal{N}(\mu, \Sigma)$ then

$$Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\mathsf{T}) \tag{7}$$

Suppose (x, y) are jointly distributed as in (6). Then,

$$\begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x \tag{8}$$

Substituting Eq. (8) in Eq. (7) we can express the distribution of x as

$$x \sim \mathcal{N}\left(\begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} \right),$$

which simplifies to $x \sim \mathcal{N}(\mu_x, \Sigma_x)$. Although this proof is shorter, it relies on (7). The previous proof has the added benefit of demonstrating that the marginal of jointly Gaussian random variables is once again Gaussian.

5 Conditional Distribution

We will now derive the conditional distribution given a joint distribution of the form

$$f(x, y) = (\text{const}) \cdot \exp\left(-\frac{1}{2} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}^{-1} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \right)$$

Using the block-UDL factorization (4b) we get

$$\begin{aligned} & f(x, y) \\ & = (\text{const}) \cdot \exp\left(-\frac{1}{2} \begin{bmatrix} \tilde{x} - \Sigma_{xy}\Sigma_y^{-1}\tilde{y} \\ \tilde{y} \end{bmatrix}^\mathsf{T} \begin{bmatrix} (\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})^{-1} & 0 \\ 0 & \Sigma_y^{-1} \end{bmatrix} \begin{bmatrix} \tilde{x} - \Sigma_{xy}\Sigma_y^{-1}\tilde{y} \\ \tilde{y} \end{bmatrix} \right) \\ & = (\text{const}) \cdot \exp\left(-\frac{1}{2}\tilde{y}^\mathsf{T}\Sigma_y^{-1}\tilde{y} \right) \cdot \exp\left[-\frac{1}{2}(\tilde{x} - \Sigma_{xy}\Sigma_y^{-1}\tilde{y})^\mathsf{T} \, (\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})^{-1}(\tilde{x} - \Sigma_{xy}\Sigma_y^{-1}\tilde{y}) \right] \end{aligned}$$

Notice that the left exponential term in the equation is the marginal distribution $f_y(y)$. Dividing by this we are left with

$$f_{x|y}(x, y) = \frac{f(x, y)}{f_y(y)} = (\text{const}) \cdot \exp\left[-\frac{1}{2}(\tilde{x} - \Sigma_{xy}\Sigma_y^{-1}\tilde{y})^\mathsf{T} \, (\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})^{-1}(\tilde{x} - \Sigma_{xy}\Sigma_y^{-1}\tilde{y}) \right]$$

From this equation we can notice that the conditional distribution is normally distributed and

$$f_{x|y}(x, y) \sim \mathcal{N}\big(\mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\big)$$

There are a few key observations that can be made about the conditional distribution

- The conditional covariance decreases (compared to the marginal distribution) after you make measurements. It is important to note that both the conditional and marginal covariance are equal only when Σ_{xy} is 0. This happens when x and y are uncorrelated.

$$\begin{gathered}\mathbf{Cov}(x) \succeq \mathbf{Cov}(x \mid y)\\ \mathbf{Cov}(x)^{-1} \preceq \mathbf{Cov}(x \mid y)^{-1}\end{gathered}$$

  The second equation shows that one uncertainty ellipsoid is contained within the other. Specifically, if we consider the confidence ellipsoid with confidence p, then $\alpha = F_{\chi_n^2}^{-1}(p)$ and

$$\begin{aligned}\Sigma_1^{-1} \preceq \Sigma_2^{-1} \qquad &\Longrightarrow \qquad x^\mathsf{T}\Sigma_1^{-1} \leq x^\mathsf{T}\Sigma_2^{-1}x\\ &\Longrightarrow \qquad x^\mathsf{T}\Sigma_2^{-1}x \leq \alpha \implies x^\mathsf{T}\Sigma_1^{-1}x \leq \alpha\\ &\Longrightarrow \qquad \left\{x \in \mathbb{R}^n \;\middle|\; x^\mathsf{T}\Sigma_2^{-1}x \leq \alpha\right\} \subseteq \left\{x \in \mathbb{R}^n \;\middle|\; x^\mathsf{T}\Sigma_1^{-1}x \leq \alpha\right\}\end{aligned}$$

  Therefore, if Σ_2^{-1} is larger than Σ_1^{-1}, then the confidence ellipsoid for Σ_2 is contained inside the confidence ellipsoid for Σ_1. Here, $\Sigma_1 = \mathbf{Cov}(x)$ and $\Sigma_2 = \mathbf{Cov}(x \mid y)$.

- The mean of the conditional distribution depends explicitly on new measurements whereas the conditional covariance does not depend on the new measurements. This means that we can know how our error will change once we receive our measurements, even before the measurements arrive. This is a unique property of Gaussian distributions and it does not hold in general for non-Gaussian distributions.