Deriving the conditional distributions of a multivariate normal distribution

166

161

We have a multivariate normal vector $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. Consider partitioning $\boldsymbol{\mu}$ and $\boldsymbol{Y}$ into

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$$

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}$$

with a similar partition of Σ into

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Then, $(\boldsymbol{y}_1 | \boldsymbol{y}_2 = \boldsymbol{a})$, the conditional distribution of the first partition given the second, is $\mathcal{N}(\overline{\boldsymbol{\mu}}, \overline{\Sigma})$, with mean

$$\overline{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \Sigma_{12} {\Sigma_{22}}^{-1} (\boldsymbol{a} - \boldsymbol{\mu}_2)$$

and covariance matrix

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12} {\Sigma_{22}}^{-1} \Sigma_{21}$$

Actually these results are provided in Wikipedia too, but I have no idea how the $\overline{\boldsymbol{\mu}}$ and $\overline{\Sigma}$ is derived. These results are crucial, since they are important statistical formula for deriving **Kalman filters**. Would anyone provide me a derivation steps of deriving $\overline{\boldsymbol{\mu}}$ and $\overline{\Sigma}$? Thank you very much!

normal-distribution conditional-probability

Share Cite Improve this question Follow

edited Jul 3, 2012 at 11:22
Macro
41.3k 10 146 149

asked Jun 16, 2012 at 18:09
Flying pig
5,869 11 32 32

38

The idea is to use the definition of conditional density $f(y_1 | y_2 = a) = \dfrac{f_{Y_1,Y_2}(y_1, a)}{f_{Y_2}(a)}$. You know that the

joint f_{Y_1,Y_2} is a bivariate normal and that the marginal f_{Y_2} is a normal then you just have to replace the values and do the unpleasant algebra. These notes might be of some help. Here is the full proof.
– user10525 Jun 16, 2012 at 18:16

1 Your second link answers the question (+1). Why not put it as an answer @Procrastinator? – gui11aume Jun 16, 2012 at 22:54

1 I hadn't realized it, but I think I was implicitly using this equation in a conditional PCA. The conditional PCA requires a transformation $\left(I - A'\left(AA'\right)^{-1}A\right)\Sigma$ that is effectively calculating the conditional covariance matrix given some choice of A. – John Jul 2, 2012 at 15:49

@Procrastinator - your approach actually requires the knowledge of the Woodbury matrix identity, and the knowledge of block-wise matrix inversion. These result in unnecessarily complicated matrix algebra.
– probabilityislogic Jul 2, 2012 at 16:17

2 @probabilityislogic Actually the result is proved in the link I provided. But it is respectable if you find it more complicated than other methods. In addition, I was not attempting to provide an optimal solution in my *comment*. Also, my comment was previous to Macro's answer (which I upvoted as you can see).
– user10525 Jul 2, 2012 at 16:25

2 Answers

Sorted by:

Highest score (default)

159

You can prove it by explicitly calculating the conditional density by brute force, as in Procrastinator's link (+1) in the comments. But, there's also a theorem that says all conditional distributions of a multivariate normal distribution are normal. Therefore, all that's left is to calculate the mean vector and covariance matrix. I remember we derived this in a time series class in college by cleverly defining a third variable and using its properties to derive the result more simply than the brute force solution in the link (as long as you're comfortable with matrix algebra). I'm going from memory but it was something like this:

Let $\mathbf{x}_1$ be the first partition and $\mathbf{x}_2$ the second. Now define $\mathbf{z} = \mathbf{x}_1 + \mathbf{A}\mathbf{x}_2$ where $\mathbf{A} = -\Sigma_{12}\Sigma_{22}^{-1}$. Now we can write

$$\begin{aligned}
\text{cov}(\mathbf{z}, \mathbf{x}_2) &= \text{cov}(\mathbf{x}_1, \mathbf{x}_2) + \text{cov}(\mathbf{A}\mathbf{x}_2, \mathbf{x}_2) \\
&= \Sigma_{12} + \mathbf{A}\text{var}(\mathbf{x}_2) \\
&= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} \\
&= 0
\end{aligned}$$

Therefore $\mathbf{z}$ and $\mathbf{x}_2$ are uncorrelated and, since they are jointly normal, they are independent. Now, clearly $E(\mathbf{z}) = \boldsymbol{\mu}_1 + \mathbf{A}\boldsymbol{\mu}_2$, therefore it follows that

$$\begin{aligned}
E(\mathbf{x}_1|\mathbf{x}_2) &= E(\mathbf{z} - \mathbf{A}\mathbf{x}_2|\mathbf{x}_2) \\
&= E(\mathbf{z}|\mathbf{x}_2) - E(\mathbf{A}\mathbf{x}_2|\mathbf{x}_2) \\
&= E(\mathbf{z}) - \mathbf{A}\mathbf{x}_2 \\
&= \boldsymbol{\mu}_1 + \mathbf{A}(\boldsymbol{\mu}_2 - \mathbf{x}_2)
\end{aligned}$$

$$= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

which proves the first part. For the covariance matrix, note that

$$\begin{aligned}
\text{var}(\mathbf{x}_1|\mathbf{x}_2) &= \text{var}(\mathbf{z} - \mathbf{A}\mathbf{x}_2|\mathbf{x}_2) \\
&= \text{var}(\mathbf{z}|\mathbf{x}_2) + \text{var}(\mathbf{A}\mathbf{x}_2|\mathbf{x}_2) - \mathbf{A}\text{cov}(\mathbf{z}, -\mathbf{x}_2) - \text{cov}(\mathbf{z}, -\mathbf{x}_2)\mathbf{A}' \\
&= \text{var}(\mathbf{z}|\mathbf{x}_2) \\
&= \text{var}(\mathbf{z})
\end{aligned}$$

Now we're almost done:

$$\begin{aligned}
\text{var}(\mathbf{x}_1|\mathbf{x}_2) = \text{var}(\mathbf{z}) &= \text{var}(\mathbf{x}_1 + \mathbf{A}\mathbf{x}_2) \\
&= \text{var}(\mathbf{x}_1) + \mathbf{A}\text{var}(\mathbf{x}_2)\mathbf{A}' + \mathbf{A}\text{cov}(\mathbf{x}_1, \mathbf{x}_2) + \text{cov}(\mathbf{x}_2, \mathbf{x}_1)\mathbf{A}' \\
&= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}
\end{aligned}$$

which proves the second part.

Note: For those not very familiar with the matrix algebra used here, this is an excellent resource.

Edit: One property used here this is not in the matrix cookbook (good catch @FlyingPig) is property 6 on the wikipedia page about covariance matrices: which is that for two random vectors $\mathbf{x}, \mathbf{y}$,

$$\text{var}(\mathbf{x} + \mathbf{y}) = \text{var}(\mathbf{x}) + \text{var}(\mathbf{y}) + \text{cov}(\mathbf{x}, \mathbf{y}) + \text{cov}(\mathbf{y}, \mathbf{x})$$

For scalars, of course, $\text{cov}(X, Y) = \text{cov}(Y, X)$ but for vectors they are different insofar as the matrices are arranged differently.

Share Cite Improve this answer Follow

edited Jun 18, 2015 at 8:18
Naetmul
103 4

answered Jun 16, 2012 at 23:29
Macro
41.3k 10 146 149

Thanks for this brilliant method! There is one matrix algebra does not seem familiar to me, where can I find the formula for opening $var(x_1 + Ax_2)$? I haven't found it on the link you sent. – Flying pig Jun 17, 2012 at 6:35

22 This is a very good answer (+1), but could be improved in terms of the ordering of the approach. We start with saying we want a linear combination $z = Cx = C_1x_1 + C_2x_2$ of the whole vector that is independent/uncorrelated with x_2. This is because we can use the fact that $p(z|x_2) = p(z)$ which means $var(z|x_2) = var(z)$ and $E(z|x_2) = E(z)$. These in turn lead to expressions for $var(C_1x_1|x_2)$ and $E(C_1x_1|x_2)$. This means we should take $C_1 = I$. Now we require $cov(z, x_2) = \Sigma_{12} + C_2\Sigma_{22} = 0$. If Σ_{22} is invertible we then have $C_2 = -\Sigma_{12}\Sigma_{22}^{-1}$. – probabilityislogic Jul 2, 2012 at 16:00

2 @probabilityislogic, I'd actually never thought about the process that resulted in choosing this linear combination but your comment makes it clear that it arises naturally, considering the constraints we want to

satisfy. +1! – Macro Jul 2, 2012 at 20:06

1 @jakeoung - it is not *proving* that $C_1 = I$, it is setting it to this value, so that we get an expression that contains the variables we want to know about. – probabilityislogic Jan 14, 2018 at 14:40

1 @jakeoung I also don't quite understand that statement. I understand in this way: If $cov(z, x_2) = 0$, then $cov(C_1^{-1} z, x_2) = C_1^{-1} cov(z, x_2) = 0$. So the value of C_1 is somehow an arbitrary scale. So we set $C_1 = I$ for simplicity. – Ken T May 5, 2018 at 16:03

22

The answer by Macro is great, but here is an even simpler way that does not require you to use any outside theorem asserting the conditional distribution. It involves writing the Mahalanobis distance in a form that separates the argument variable for the conditioning statement, and then factorising the normal density accordingly.

Rewriting the Mahalanobis distance for a conditional vector: This derivation uses a matrix inversion formula that uses the Schur complement $\boldsymbol{\Sigma}_* \equiv \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. We first use the blockwise inversion formula to write the inverse-variance matrix as:

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^* & \boldsymbol{\Sigma}_{12}^* \\ \boldsymbol{\Sigma}_{21}^* & \boldsymbol{\Sigma}_{22}^* \end{bmatrix},$$

where:

$$\boldsymbol{\Sigma}_{11}^* = \boldsymbol{\Sigma}_*^{-1} \qquad \boldsymbol{\Sigma}_{12}^* = -\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1},$$

$$\boldsymbol{\Sigma}_{21}^* = -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_*^{-1} \qquad \boldsymbol{\Sigma}_{22}^* = \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}.$$

Using this formula we can now write the Mahalanobis distance as:

$$\begin{aligned}
&(\boldsymbol{y}-\boldsymbol{\mu})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \\
&= \begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{y}_2 - \boldsymbol{\mu}_2 \end{bmatrix}^{\text{T}} \begin{bmatrix} \boldsymbol{\Sigma}_{11}^* & \boldsymbol{\Sigma}_{12}^* \\ \boldsymbol{\Sigma}_{21}^* & \boldsymbol{\Sigma}_{22}^* \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{y}_2 - \boldsymbol{\mu}_2 \end{bmatrix} \\
&= (\boldsymbol{y}_1 - \boldsymbol{\mu}_1)^{\text{T}}\boldsymbol{\Sigma}_{11}^*(\boldsymbol{y}_1 - \boldsymbol{\mu}_1) + (\boldsymbol{y}_1 - \boldsymbol{\mu}_1)^{\text{T}}\boldsymbol{\Sigma}_{12}^*(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\
&\quad + (\boldsymbol{y}_2 - \boldsymbol{\mu}_2)^{\text{T}}\boldsymbol{\Sigma}_{21}^*(\boldsymbol{y}_1 - \boldsymbol{\mu}_1) + (\boldsymbol{y}_2 - \boldsymbol{\mu}_2)^{\text{T}}\boldsymbol{\Sigma}_{22}^*(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\
&= (\boldsymbol{y}_1 - \boldsymbol{\mu}_1)^{\text{T}}\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1) - (\boldsymbol{y}_1 - \boldsymbol{\mu}_1)^{\text{T}}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\
&\quad - (\boldsymbol{y}_2 - \boldsymbol{\mu}_2)^{\text{T}}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1) + (\boldsymbol{y}_2 - \boldsymbol{\mu}_2)^{\text{T}}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\
&\quad + (\boldsymbol{y}_2 - \boldsymbol{\mu}_2)^{\text{T}}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \\
&= (\boldsymbol{y}_1 - (\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2)))^{\text{T}}\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{y}_1 - (\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2)))
\end{aligned}$$

$$+ (y_2 - \mu_2)^{\text{T}} \Sigma_{22}^{-1} (y_2 - \mu_2)$$

$$= (y_1 - \mu_*)^{\text{T}} \Sigma_*^{-1} (y_1 - \mu_*) + (y_2 - \mu_2)^{\text{T}} \Sigma_{22}^{-1} (y_2 - \mu_2),$$

where $\mu_* \equiv \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2)$ is the **conditional mean vector**. Note that this result is a general result that does not assume normality of the random vectors involved in the decomposition. It gives a useful way of decomposing the Mahalanobis distance so that it consists of a sum of quadratic forms on the marginal and conditional parts. In the conditional part the conditioning vector y_2 is absorbed into the mean vector and variance matrix. To clarify the form, we repeat the equation with labelling of terms:

$$(y - \mu)^{\text{T}} \Sigma^{-1} (y - \mu) = \underbrace{(y_1 - \mu_*)^{\text{T}} \Sigma_*^{-1} (y_1 - \mu_*)}_{\text{Conditional Part}} + \underbrace{(y_2 - \mu_2)^{\text{T}} \Sigma_{22}^{-1} (y_2 - \mu_2)}_{\text{Marginal Part}}.$$

Deriving the conditional distribution: Now that we have the above form for the Mahalanobis distance, the rest is easy. We have:

$$\begin{aligned} p(y_1 | y_2, \mu, \Sigma) &\overset{y_1}{\propto} p(y_1, y_2 | \mu, \Sigma) \\ &= \text{N}(y | \mu, \Sigma) \\ &\overset{y_1}{\propto} \exp \Big(-\frac{1}{2} (y - \mu)^{\text{T}} \Sigma^{-1} (y - \mu) \Big) \\ &\overset{y_1}{\propto} \exp \Big(-\frac{1}{2} (y_1 - \mu_*)^{\text{T}} \Sigma_*^{-1} (y_1 - \mu_*) \Big) \\ &\overset{y_1}{\propto} \text{N}(y_1 | \mu_*, \Sigma_*). \end{aligned}$$

This establishes that the conditional distribution is also multivariate normal, with the specified conditional mean vector and conditional variance matrix.

Share Cite Improve this answer Follow

edited Nov 23, 2021 at 19:10

answered Feb 15, 2019 at 12:01

Ben
98.4k 3 175 404

Hi Ben. I am sorry for another question. Will the above marginal distribution hold if we don't assume normal distribution for y_1 and y_2. Then what's the conditional distribution for y_1 conditional y_2 without normal distribution? Or, is it possible to calculate the expectation and variance of y_1 conditional y_2 without normal distribution following your setup without normality assumption. Deeply appreciate for your help! Thanks – Charles Chou Nov 24, 2021 at 16:48

@CharlesChou: No, the moments μ_* and Σ_* will not generally hold outside the normal distribution. (Also, note that the above is a conditional distribution, not a marginal distribution.) – Ben Nov 24, 2021 at 20:24