2012-09-14

Technical Note on Manipulating Multivariate Gaussian Distributions

John C. Macdonald

Technical Note on Manipulating Multivariate Gaussian Distributions

John Macdonald

In this technical note we present some derivations treating conditional, marginal, and joint distributions for Gaussian random vectors. First we present the intermediate steps that allow the conditional distribution $\mathrm{p}\left(\mathbf{x}_1|\mathbf{x}_2\right)$ to be rewritten in a form that admits easier multiplication with the marginal distribution $\mathrm{p}\left(\mathbf{x}_2\right)$. As a corollary we make some observations about the relationship that exists between the information matrices for the conditional, marginal, and joint distributions. We conclude with a derivation showing how to recover the joint covariance matrix of $\mathrm{p}\left(\mathbf{x}_1, \mathbf{x}_2\right)$ without matrix inversion to allow information updated in the marginal to be distributed to the remaining variables in the joint.

Most of the motivation for these derivations derives from an apparently unpublished note written by Paul Newman and John Leonard of MIT in November 2002. We correct some small errors in that note and extend it to include the closed form expressions for recovering the optimized joint mean and covariance without inverting the optimized joint information matrix.

I. Rewriting the Conditional Distribution

Let a D-dimensional random vector $\mathbf{x}$ of jointly Gaussian variables be partitioned into two, disjoint sub-vectors such that $\mathbf{x} = \left[\mathbf{x}_1^\top, \mathbf{x}_2^\top\right]^\top$, where $\mathbf{x}_1$ is dimension D_1 and $\mathbf{x}_2$ is dimension D_2. Then the joint distribution $\mathrm{p}(\mathbf{x})$, with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, is partitioned such that

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}.$$

Typical textbook derivations (e.g. [1], Chapter 2.3.1) define the conditional distribution $\mathrm{p}\left(\mathbf{x}_1|\mathbf{x}_2\right)$ such that $\log\left(\mathrm{p}\left(\mathbf{x}_1|\mathbf{x}_2\right)\right)$ is proportional to

$$\left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right)^\top \boldsymbol{\Sigma}_{1|2}^{-1} \left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right), \tag{1}$$

where $\boldsymbol{\mu}_{1|2}$ and $\boldsymbol{\Sigma}_{1|2}$ are given by

$$\boldsymbol{\mu}_{1|2} \triangleq \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right), \tag{2}$$

$$\boldsymbol{\Sigma}_{1|2} \triangleq \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{3}$$

To simplify notation in the sequel we define

$$\mathbf{K} \triangleq \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}. \tag{4}$$

There are a few undesirable aspects of expressing the conditional distribution $\mathrm{p}\left(\mathbf{x}_1|\mathbf{x}_2\right)$ using (1). First, the mean vector $\boldsymbol{\mu}_{1|2}$ has dimension D_1. The marginal distribution $\mathrm{p}\left(\mathbf{x}_2\right)$ has a mean $\boldsymbol{\mu}_2$ of dimension D_2. To recover the joint distribution $\mathrm{p}\left(\mathbf{x}\right) = \mathrm{p}\left(\mathbf{x}_1|\mathbf{x}_2\right)\mathrm{p}\left(\mathbf{x}_2\right)$ would require that we sum exponents with different dimensions. It is also unattractive to leave the conditional distribution's functional dependence on $\mathbf{x}_2$ buried in the conditional mean.

Define the following variables:

$$\mathbf{M} \triangleq \begin{bmatrix}\mathbf{I}_{D_1} & -\mathbf{K}\end{bmatrix}, \quad \mathbf{g} \triangleq \left(\mathbf{x} - \boldsymbol{\mu}\right) = \begin{bmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{bmatrix},$$

where $\mathbf{I}_{D_1}$ is a D_1 x D_1 identity matrix. Note that $\mathbf{M}$ is size D_1 x D. Now observe that

$$\begin{aligned}\mathbf{Mg} &= \mathbf{x}_1 - \boldsymbol{\mu}_1 - \mathbf{K}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) \\ &= \mathbf{x}_1 - \boldsymbol{\mu}_{1|2}.\end{aligned} \tag{5}$$

Substituting Equation (5) into Equation (1) gives

$$\begin{aligned}\left(\mathbf{Mg}\right)^\top \boldsymbol{\Sigma}_{1|2}^{-1}\left(\mathbf{Mg}\right) &= \mathbf{g}^\top\mathbf{M}^\top\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{Mg} \\ &= \left(\mathbf{x} - \boldsymbol{\mu}\right)^\top \mathbf{A}\left(\mathbf{x} - \boldsymbol{\mu}\right),\end{aligned} \tag{6}$$

where $\mathbf{A} \triangleq \mathbf{M}^\top\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{M}$ is the information matrix associated with the conditional distribution. We can expand this expression for $\mathbf{A}$ to find that

$$\mathbf{A} = \begin{bmatrix}\boldsymbol{\Sigma}_{1|2}^{-1} & -\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K} \\ -\mathbf{K}^\top\boldsymbol{\Sigma}_{1|2}^{-1} & \mathbf{K}^\top\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}\end{bmatrix}. \tag{7}$$

Equation (6) gives an expression for the conditional distribution where its functional dependence on both $\mathbf{x}_1$ and $\mathbf{x}_2$ is in a more standard form. We can similarly rewrite the marginal $\mathrm{p}\left(\mathbf{x}_2\right)$ such that $\log\left(\mathrm{p}\left(\mathbf{x}_2\right)\right)$ is proportional to

$$\left(\mathbf{x} - \mathbf{b}\right)^\top \mathbf{B}\left(\mathbf{x} - \mathbf{b}\right),$$

where the D-dimensional vector $\mathbf{b}$ is defined as

$$\mathbf{b} \triangleq \begin{bmatrix}\mathbf{0} \\ \boldsymbol{\mu}_2\end{bmatrix},$$

and

$$\mathbf{B} \triangleq \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1}\end{bmatrix}. \tag{8}$$

II. A Short Observation on Information Matrices

A nice piece of intuition arises from writing the conditional and marginal distributions in the manner just described. The block partitioned information matrix for the joint distribution can be written [1] as

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix}\boldsymbol{\Sigma}_{1|2}^{-1} & -\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K} \\ -\mathbf{K}^\top\boldsymbol{\Sigma}_{1|2}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \mathbf{K}^\top\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}\end{bmatrix}. \tag{9}$$

Comparing Equations (7) and (8) to Equation (9), it is clear that conditioning $\mathbf{x}_1$ on $\mathbf{x}_2$ amounts to removing the marginal information associated with $\mathbf{x}_2$; i.e.

$$\mathbf{A} = \boldsymbol{\Sigma}^{-1} - \mathbf{B}.$$

Conversely, when we recover a joint distribution, perhaps with an improved marginal belief about $\mathbf{x}_2$, we are simply adding that information back in:

$$\boldsymbol{\Sigma}^{-1} = \mathbf{A} + \mathbf{B}.$$

III. Deriving the Updated Joint Covariance Without Matrix Inversion

Next we present the steps that lead to a closed form expression for the joint covariance that has been optimized to include new information from the marginal distribution. Let $\mathrm{p}(\check{\mathbf{x}}_2)$, with mean $\check{\boldsymbol{\mu}}_2$ and information matrix $\check{\boldsymbol{\Sigma}}_{22}^{-1}$, represent our updated belief about the marginal states after incorporating information from a measurement. We also define

$$\check{\mathbf{b}} \triangleq \begin{bmatrix} \mathbf{0} \\ \check{\boldsymbol{\mu}}_2 \end{bmatrix}$$
$$\check{\mathbf{B}} \triangleq \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \check{\boldsymbol{\Sigma}}_{22}^{-1} \end{bmatrix}$$

To recover the optimized joint distribution $\mathrm{p}(\mathring{\mathbf{x}})$, we must perform the multiplication $\mathrm{p}(\mathbf{x}_1|\mathbf{x}_2)\mathrm{p}(\check{\mathbf{x}}_2)$. The product leads to

$$\begin{aligned}\log\left(\mathrm{p}\left(\mathring{\mathbf{x}}\right)\right) &\propto (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) + (\mathbf{x}-\check{\mathbf{b}})^\top \check{\mathbf{B}}(\mathbf{x}-\check{\mathbf{b}}) \\ &\propto (\mathbf{x}-\mathring{\boldsymbol{\mu}})^\top \mathring{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\mathring{\boldsymbol{\mu}}),\end{aligned}$$

where the optimized joint covariance and mean are

$$\begin{aligned}\mathring{\boldsymbol{\Sigma}} &\triangleq \left(\mathbf{A}+\check{\mathbf{B}}\right)^{-1} \\ &= \begin{bmatrix} \boldsymbol{\Sigma}_{1|2}^{-1} & -\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K} \\ -\mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1} & \check{\boldsymbol{\Sigma}}_{22}^{-1} + \mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K} \end{bmatrix}^{-1} \qquad (10) \\ \mathring{\boldsymbol{\mu}} &\triangleq \mathring{\boldsymbol{\Sigma}}\left(\mathbf{A}\boldsymbol{\mu} + \check{\mathbf{B}}\check{\mathbf{b}}\right) \qquad (11)\end{aligned}$$

We will use the following identity (Chapter 0.7.3 of [2]) for the inverse of a block partitioned matrix:

$$\begin{bmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Delta}_1 & -\mathbf{W}^{-1}\mathbf{X}\boldsymbol{\Delta}_2 \\ -\boldsymbol{\Delta}_2\mathbf{Y}\mathbf{W}^{-1} & \boldsymbol{\Delta}_2 \end{bmatrix}, \qquad (12)$$

where

$$\begin{aligned}\boldsymbol{\Delta}_1 &\triangleq \left(\mathbf{W}-\mathbf{X}\mathbf{Z}^{-1}\mathbf{Y}\right)^{-1} \\ &= \mathbf{W}^{-1} + \mathbf{W}^{-1}\mathbf{X}\boldsymbol{\Delta}_2\mathbf{Y}\mathbf{W}^{-1}, \qquad (13)\end{aligned}$$

and

$$\boldsymbol{\Delta}_2 \triangleq \left(\mathbf{Z}-\mathbf{Y}\mathbf{W}^{-1}\mathbf{X}\right)^{-1}. \qquad (14)$$

Comparing Equation (10) and Equation (12) we have

$$\begin{aligned}\mathbf{W} &= \boldsymbol{\Sigma}_{1|2}^{-1}, \\ \mathbf{X} &= -\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}, \\ \mathbf{Y} &= -\mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1}, \\ \mathbf{Z} &= \check{\boldsymbol{\Sigma}}_{22}^{-1} + \mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}. \qquad (15)\end{aligned}$$

Our goal is to recover the optimized joint covariance matrix $\mathring{\boldsymbol{\Sigma}}$ without matrix inversion. We begin by finding the bottom right block $\mathring{\boldsymbol{\Sigma}}_{22}$ of the new joint covariance; this corresponds to the bottom right block $\boldsymbol{\Delta}_2$ of Equation (12). Making the appropriate substitutions into Equation (14) leads to

$$\begin{aligned}\mathring{\boldsymbol{\Sigma}}_{22} &= \boldsymbol{\Delta}_2 \\ &= \left[\mathbf{Z} - \left(-\mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1}\right)\left(\boldsymbol{\Sigma}_{1|2}\right)\left(-\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}\right)\right]^{-1} \\ &= \left[\mathbf{Z} - \mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}\right]^{-1} \\ &= \left[\check{\boldsymbol{\Sigma}}_{22}^{-1} + \mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K} - \mathbf{K}^\top \boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}\right]^{-1} \\ &= \check{\boldsymbol{\Sigma}}_{22}. \qquad (16)\end{aligned}$$

This agrees nicely with intuition. The optimized uncertainty for the variables $\mathring{\mathbf{x}}_2$ in the joint distribution is the same as the uncertainty in the updated marginal distribution.

Now consider the upper left block $\mathring{\boldsymbol{\Sigma}}_{11}$; this corresponds to $\boldsymbol{\Delta}_1$ in Equation (12). We first recall from Equation (3) that $\mathbf{W} = \boldsymbol{\Sigma}_{1|2}^{-1}$ can be rewritten such that

$$\begin{aligned}\mathbf{W}^{-1} &= \boldsymbol{\Sigma}_{1|2} \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{11} - \mathbf{K}\boldsymbol{\Sigma}_{22}\mathbf{K}^\top. \qquad (17)\end{aligned}$$

Then observe that

$$\begin{aligned}\mathbf{W}^{-1}\mathbf{X} &= \left(\boldsymbol{\Sigma}_{1|2}\right)\left(-\boldsymbol{\Sigma}_{1|2}^{-1}\mathbf{K}\right) \\ &= -\mathbf{K} \\ &= \left(\mathbf{Y}\mathbf{W}^{-1}\right)^\top. \qquad (18)\end{aligned}$$

Using Equations (19), (17), and (18) allows us to simplify Equation (13) such that

$$\begin{aligned}\mathring{\boldsymbol{\Sigma}}_{11} &= \boldsymbol{\Delta}_1 \\ &= \mathbf{W}^{-1} + \mathbf{W}^{-1}\mathbf{X}\boldsymbol{\Delta}_2\mathbf{Y}\mathbf{W}^{-1} \\ &= \boldsymbol{\Sigma}_{11} - \mathbf{K}\boldsymbol{\Sigma}_{22}\mathbf{K}^\top + \mathbf{K}\check{\boldsymbol{\Sigma}}_{22}\mathbf{K}^\top \\ &= \boldsymbol{\Sigma}_{11} - \mathbf{K}\left(\boldsymbol{\Sigma}_{22} - \check{\boldsymbol{\Sigma}}_{22}\right)\mathbf{K}^\top. \qquad (19)\end{aligned}$$

Equation (19) also has an intuitive explanation. A decrease in the uncertainty of the conditioned variables (i.e. $\mathbf{x}_1$) is proportional to a decrease in the uncertainty of the conditioning variables (i.e. $\mathbf{x}_2$).

Finally, we want to find the updated cross-covariance terms $\mathring{\boldsymbol{\Sigma}}_{12} = \left(\mathring{\boldsymbol{\Sigma}}_{21}\right)^\top$. From Equation (12) we get

$$\mathring{\boldsymbol{\Sigma}}_{12} = -\mathbf{W}^{-1}\mathbf{X}\boldsymbol{\Delta}_2$$

which from the foregoing derivations can be readily expressed as

$$= \mathbf{K}\check{\boldsymbol{\Sigma}}_{22}. \qquad (20)$$

To restate it in one place, the joint covariance after incorporating the updated marginal information is given by

$$\begin{aligned}\mathring{\boldsymbol{\Sigma}} &= \begin{bmatrix} \mathring{\boldsymbol{\Sigma}}_{11} & \mathring{\boldsymbol{\Sigma}}_{12} \\ \mathring{\boldsymbol{\Sigma}}_{21} & \mathring{\boldsymbol{\Sigma}}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Sigma}_{11} - \mathbf{K}\left(\boldsymbol{\Sigma}_{22} - \check{\boldsymbol{\Sigma}}_{22}\right)\mathbf{K}^\top & \mathbf{K}\check{\boldsymbol{\Sigma}}_{22} \\ \hat{\boldsymbol{\Sigma}}_{22}\mathbf{K}^\top & \check{\boldsymbol{\Sigma}}_{22} \end{bmatrix}. \qquad (21)\end{aligned}$$

These terms are all available when the optimized joint information matrix is formed, therefore the optimized covariance matrix can be calculated directly without inverting the information matrix. Finding the new joint mean follows similar steps which we omit here in the interest of brevity. The result is suggested by Equation (21):

$$\overset{\star}{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{\mu}_1 - \mathbf{K}\left(\boldsymbol{\mu}_2 - \check{\boldsymbol{\mu}}_2\right) \\ \check{\boldsymbol{\mu}}_2 \end{bmatrix}. \tag{22}$$

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2009.