The Multivariate Gaussian Distribution

David S. Rosenberg

Notes from DS-GA 1002: Probability_3 Section 2.4: Gaussian Random Vectors (pp. 23-26).

[TO DO: need to put something about inverses of block matrices; Also need to talk about iterated expectations]

See also: Murphy p. 113 Section 4.3.1.

1 One-Dimensional Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

2 Multivariate Gaussian Density

A random vector $x \in \mathbf{R}^d$ has a d**-dimensional multivariate Gaussian distribution** with mean $\mu \in \mathbf{R}^d$ and covariance matrix $\Sigma \in \mathbf{R}^{d\times d}$ if its density is given by

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

where $|\Sigma|$ denotes the determinant of Σ. Note that this expression requires that the covariance matrix Σ be invertible[1]. Sometimes we will rewrite the factor in front of the $\exp(\cdot)$ as $|2\pi\Sigma|^{-1/2}$, which follows from basic facts about determinants.

[1] We **can** have a d-dimensional Gaussian distribution with a non-invertible Σ, but such a distribution will not have a density on $\mathbf{R}^n$, and that case will not be of interest here.

Exercise 1. There are at least 2 claims implicit in this definition. First, that the expression given is, in fact, a density(i.e. it's non-negative and integrates to 1). Second, the density corresponds to a distribution with mean μ and covariance Σ, as claimed.

3 Recognizing a Gaussian Density

If we come across a density function of the form $p(x) \propto e^{-q(x)/2}$, where $q(x)$ is a positive definite quadratic function, then $p(x)$ is the density for a Gaussian distribution. More precisely, we have the following theorem:

Theorem 2. *Consider the quadratic function $q(x) = x^T \Lambda x - 2b^T x + c$, for any **symmetric positive definite** $A \in \mathbf{R}^{d\times d}$, any $b \in \mathbf{R}^d$, and $c \in \mathbf{R}$. If $p(x)$ is a density function with*

$$p(x) \propto e^{-q(x)/2},$$

then $p(x)$ is a multivariate Gaussian density with mean $\Lambda^{-1}b$ and covariance Λ^{-1}. That is,

$$p(x) = \frac{|\Lambda|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x - \Lambda^{-1}b)^T \Lambda (x - \Lambda^{-1}b)\right).$$

Note: The inverse of the covariance matrix is called the **precision matrix**. Precision matrices of multivariate Gaussians have some interesting properties. [explain that this is the Gaussian density in "information form" or "canonical form" c.f. Murphy p. 117).]

Proof. Completing the square, we have

$$\begin{aligned} q(x) &= x^T \Lambda x - 2b^T x + c \\ &= \left(x - \Lambda^{-1}b\right)^T \Lambda (x - \Lambda^{-1}b) - b^T \Lambda^{-1} b + c. \end{aligned}$$

Since the last two terms are independent of x, when we exponentiate $q(x)$, they can be absorbed into the constant of proportionality. That is,

$$\begin{aligned} e^{-q(x)/2} &= \exp\left[-\frac{1}{2}\left(x - \Lambda^{-1}b\right)^T \Lambda (x - \Lambda^{-1}b)\right] \exp\left(-\frac{1}{2}\left[-b^T \Lambda^{-1} b + c\right]\right) \\ &\propto \exp\left[-\frac{1}{2}\left(x - \Lambda^{-1}b\right)^T \Lambda (x - \Lambda^{-1}b)\right] \end{aligned}$$

Now recall that the density function for the multivariate Gaussian density $\mathcal{N}(\mu, \Sigma)$ is

$$\phi(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

Thus we see that $p(x)$ must also be a Gaussian density with covariance $\Sigma = \Lambda^{-1}$ and mean $\Lambda^{-1}b$. $\square$

4 Conditional Distributions (Bishop Section 2.3.1)

Let $x \in \mathbf{R}^d$ have a Gaussian distribution: $x \sim \mathcal{N}(\mu, \Sigma)$. Let's partition the random variables in x into two pieces:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where $x_1 \in \mathbf{R}^{d_1}, x_2 \in \mathbf{R}^{d_2}$ and $d = d_1 + d_2$. Similarly, we'll partition the mean vector, the covariance matrix, and the precision matrix as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix},$$

where $\mu_1 \in \mathbf{R}^{d_1}$, $\Sigma_{12} \in \mathbf{R}^{d_1 \times d_2}$, $\Lambda_{12} \in \mathbf{R}^{d_1 \times d_2}$, etc. Note that by the symmetry of the covariance matrix Σ, we have $\Sigma_{12} = \Sigma_{21}^T$.

When $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ has a Gaussian distribution, we say that x_1 and x_2 are **jointly Gaussian**. Can we conclude anything about the marginal distributions of x_1 and x_2? Indeed, the following theorem states that they are individually Gaussian:

Theorem 3. . *Let x, μ, and Σ be as defined above. Then the marginal distributions of x_1 and x_2 are each Gaussian, with*

$$\begin{aligned} x_1 &\sim \mathcal{N}(\mu_1, \Sigma_1) \\ x_2 &\sim \mathcal{N}(\mu_2, \Sigma_2). \end{aligned}$$

Proof. (See Bishop Section 2.3.2, p. 88) This can be done by showing that the marginal density $p(x_1) = \int p(x_1, x_2)\, dx_2$ has the form claimed, and similarly for x_2. $\square$

So when x_1 and x_2 are jointly Gaussian, we know that x_1 and x_2 are also marginally Gaussian. It turns out that the conditional distributions $x_1 \mid x_2$ and $x_2 \mid x_1$ are also Gaussian:

Theorem 4. *Let x, μ, and Σ be as defined above. Assume that Σ_{22} is positive definite*[2]*. Then the distribution of x_1 given x_2 is multivariate normal. More specifically,*

$$x_1 \mid x_2 \sim \mathcal{N}\left(\mu_{1|2}, \Sigma_{1|2}\right),$$

where

$$\begin{aligned}\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.\end{aligned}$$

Proof. (See Bishop Section 2.3.1, p. 85) □

Example. Consider a standard regression framework in which we are building a predictive model for $x_1 \in \mathbf{R}$ given $x_2 \in \mathbf{R}^d$. Recall that if we are using a square loss, then the Bayes optimal prediction function is $f^*(x_2) = \mathbb{E}\left[x_1 \mid x_2\right]$. If we assume that x_1 and x_2 are jointly Gaussian with a positive definite covariance matrix, then Theorem 4 gives us tells us that

$$\mathbb{E}\left[x_1 \mid x_2\right] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2).$$

Of course, in practice we don't know μ and Σ. Nevertheless, what's interesting is that the Bayes optimal prediction function is an affine function of x_2 (i.e. a linear function plus a constant). Thus if we think that our input vector x_2 and our response variable x_1 are jointly Gaussian, there's no reason to go beyond a hypothesis space of affine functions of x_2. In other words, linear regression is all we need.

5 Joint Distribution from Marginal + Conditional

In Section 4, we found that if x_1 and x_2 are jointly Gaussian, then x_2 is marginally Gaussian and the conditional distribution $x_1 \mid x_2$ was also Gaussian, where the mean is a linear function of x_2. The following theorem shows that we can **we can go in the reverse direction as well**.

[2] In fact, this is implied by our assumption that Σ is positive definite.

Theorem. *Suppose* $x_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ *and* $x_2 \mid x_1 \sim \mathcal{N}\left(Ax_1 + b, \Sigma_{2|1}\right)$*, for some* $\mu_1 \in \mathbf{R}^{d_1}$*,* $\Sigma_1 \in \mathbf{R}^{d_1 \times d_1}$*,* $A \in \mathbf{R}^{d_2 \times d_1}$*, and* $\Sigma_{2|1} \in \mathbf{R}^{d_2 \times d_2}$*. Then* x_1 *and* x_2 *are jointly Gaussian with*

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ A\mu_1 + b \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_1 A^T \\ A\Sigma_1 & \Sigma_{2|1} + A\Sigma_1 A^T \end{pmatrix}\right).$$

We'll prove this with two steps. First, we'll show that the mean and variance of x take the form claimed above. Then, we'll write down the joint density $p(x_1, x_2) = p(x_1)p(x_2 \mid x_1)$ and show that it's proportional to $e^{-q(x)/2}$ for an appropriate quadratic $q(x)$. The result then follows from 2.

Proof. We're given that $\mathbb{E}x_1 = \mu_1$. For the other part of the mean vector, note that

$$\begin{aligned} \mathbb{E}x_2 &= \mathbb{E}\mathbb{E}\left[x_2 \mid x_1\right] \\ &= \mathbb{E}\left(Ax_1 + b\right) = A\mu_1 + b, \end{aligned}$$

which explains the lower entry in the mean.

We are given that the marginal covariance of x_1 is Σ_1. That is,

$$\mathbb{E}\left(x_1 - \mu_1\right)\left(x_1 - \mu_1\right)^T = \Sigma_1.$$

We're also given the conditional covariance of x_2:

$$\mathbb{E}\left[\left(x_2 - Ax_1 - b\right)\left(x_2 - Ax_1 - b\right)^T \mid x_1\right] = \Sigma_{2|1}.$$

We'll know try to express $\mathrm{Cov}(x_2)$ in terms of these expressions above. For convenience, we'll introduce the random variable $m_{2|1} = Ax_1 + b$. (It's random because it depends on x_1.) Note that $\mathbb{E}m_{2|1} = \mathbb{E}x_2 = A\mu_1 + b$. So

$$\begin{aligned} \mathrm{Cov}(x_2) &= \mathbb{E}\left(x_2 - \mathbb{E}x_2\right)\left(x_2 - \mathbb{E}x_2\right)^T \text{ (by definition)} \\ &= \mathbb{E}\mathbb{E}\left[\left(x_2 - \mathbb{E}x_2\right)\left(x_2 - \mathbb{E}x_2\right)^T \mid x_1\right] \text{ (law of iterated expectations)} \\ &= \mathbb{E}\mathbb{E}\left[\left(x_2 \underbrace{- m_{2|1} + m_{2|1}}_{=0} - \mathbb{E}x_2\right)\left(x_2 \underbrace{- m_{2|1} + m_{2|1}}_{=0} - \mathbb{E}x_2\right)^T \mid x_1\right] \\ &= \mathbb{E}\mathbb{E}\left[\left(\left(x_2 - m_{2|1}\right) + \left(m_{2|1} - \mathbb{E}x_2\right)\right)\left(\left(x_2 - m_{2|1}\right) + \left(m_{2|1} - \mathbb{E}x_2\right)\right)^T \mid x_1\right] \\ &= U + 2V + W, \end{aligned}$$

where we've multiplied out the parenthesized terms. The terms are as follows:

$$
\begin{aligned}
U &= \mathbb{E}\mathbb{E}\left[\left(x_2 - m_{2|1}\right)\left(x_2 - m_{2|1}\right)^T \mid x_1\right] \\
&= \Sigma_{2|1}
\end{aligned}
$$

The cross-term turns out to be zero:

$$
\begin{aligned}
V &= \mathbb{E}\mathbb{E}\left[\left(x_2 - m_{2|1}\right)\left(m_{2|1} - \mathbb{E}x_2\right)^T \mid x_1\right] \\
& \quad \mathbb{E}\mathbb{E}\left[\left(x_2 - Ax_1 - b\right)\left(Ax_1 + b - A\mu_1 - b\right)^T \mid x_1\right] \\
&= \mathbb{E}\left[\underbrace{\mathbb{E}\left[\left(x_2 - Ax_1 + b\right) \mid x_1\right]}_{=0}\left(Ax_1 + b - A\mu_1 - b\right)^T\right] \\
&= 0,
\end{aligned}
$$

where in the second to last step we used the fact that $\mathbb{E}\left[f(x)g(x,y) \mid x\right] = f(x)\mathbb{E}\left[g(x,y) \mid x\right]$. This same identity is used a couple more times below. Finally the last term is

$$
\begin{aligned}
W &= \mathbb{E}\mathbb{E}\left[\left(m_{2|1} - \mathbb{E}m_{2|1}\right)\left(m_{2|1} - \mathbb{E}m_{2|1}\right)^T \mid x_1\right] \\
&= \mathbb{E}\mathbb{E}\left[\left(Ax_1 - A\mu_1\right)\left(Ax_1 - A\mu_1\right)^T \mid x_1\right] \\
&= \mathbb{E}\left[\left(Ax_1 - A\mu_1\right)\left(Ax_1 - A\mu_1\right)^T\right] \\
&= A\left[\mathbb{E}\left(x_1 - \mu_1\right)\left(x_1 - \mu_1\right)^T\right]A^T \\
&= A\Sigma_1 A^T
\end{aligned}
$$

So

$$
\mathrm{Cov}(x_2) = \Sigma_{2|1} + A\Sigma_1 A^T,
$$

The top-right cross-covariance submatrix can be computed as follows:

$$
\begin{aligned}
\mathbb{E}\left(x_1 - \mu_1\right)\left(x_2 - A\mu_1 - b\right)^T &= \mathbb{E}\mathbb{E}\left[\left(x_1 - \mu_1\right)\left(x_2 - A\mu_1 - b\right)^T \mid x_1\right] \\
&= \mathbb{E}\left[\left(x_1 - \mu_1\right)\mathbb{E}\left[\left(x_2 - A\mu_1 - b\right)^T \mid x_1\right]\right] \\
&= \mathbb{E}\left[\left(x_1 - \mu_1\right)\left(Ax_1 + b - A\mu_1 - b\right)^T\right] \\
&= \mathbb{E}\left[\left(x_1 - \mu_1\right)\left(x_1 - \mu_1\right)^T\right]A^T \\
&= \Sigma_1 A^T.
\end{aligned}
$$

Finally, the bottom left cross-covariance matrix is just the transpose of the top right.

So far we have shown that the $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ has the mean and covariance specified in the theorem statement. We now show that the joint density is indeed Gaussian:

$$\begin{aligned}
p(x_1, x_2) &= p(x_1)p(x_2 \mid x_1) \\
&= \mathcal{N}\left(x_1 \mid \mu_1, \Sigma_1\right)\mathcal{N}\left(x_2 \mid Ax_1 + b, \Sigma_{2|1}\right) \\
&\propto \exp\left(-\frac{1}{2}(x_1 - \mu_1)^T\Sigma_1^{-1}(x_1 - \mu_1)\right) \\
&\quad \times \exp\left(-\frac{1}{2}(x_2 - Ax_1 - b)^T\Sigma_{2|1}^{-1}(x_2 - Ax_1 - b)\right) \\
&= e^{-q(x)/2},
\end{aligned}$$

where

$$q(x) = (x_1 - \mu_1)^T\Sigma_1^{-1}(x_1 - \mu_1) + (x_2 - Ax_1 - b)^T\Sigma_{2|1}^{-1}(x_2 - Ax_1 - b).$$

To apply Theorem 2, we need to make sure we can write the quadratic terms of $q(x)$ as $x^T Mx$, where M is symmetric positive definite. We'll separate the quadratic terms in $q(x)$ and write **l.o.t. for "lower order terms"**, which includes linear terms of the form $b^T x$ and constants:

$$\begin{aligned}
q(x) &= -\frac{1}{2}\left[x_2^T\Sigma_{2|1}^{-1}x_2 - 2x_1^T A^T\Sigma_{2|1}^{-1}x_2 + x_1^T\left(\Sigma_1^{-1} + A^T\Sigma_{2|1}^{-1}A\right)x_1\right] + \text{l.o.t.} \\
&= -\frac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} \left(\Sigma_1^{-1} + A^T\Sigma_{2|1}^{-1}A\right) & -A^T\Sigma_{2|1}^{-1} \\ -\Sigma_{2|1}^{-1}A & \Sigma_{2|1}^{-1} \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \text{l.o.t.}
\end{aligned}$$

Let M be that matrix in the middle. We only need to show that M is positive definite. From the Schur complement condition, M is positive definite if and only if both $\Sigma_{2|1}^{-1}$ and $M/\Sigma_{2|1}^{-1}$ are positive definite, where

$$\begin{aligned}
M/\Sigma_{2|1}^{-1} &= \left(\Sigma_1^{-1} + A^T\Sigma_{2|1}^{-1}A\right) - \left(-A^T\Sigma_{2|1}^{-1}\Sigma_{2|1}\left(-\Sigma_{2|1}^{-1}A\right)\right) \\
&= \Sigma_1^{-1}.
\end{aligned}$$

Since $\Sigma_{2|1}^{-1}$ and Σ_1^{-1} are both inverses of covariance matrices (by assumption), they are each positive definite. Thus M must be positive definite.

Thus $p(x) \propto e^{-q(x)/2}$, where $q(x)$ has the form requried by Theorem 2. We conclude that $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ is jointly Gaussian. We have also shown that the marginal means and covariances, as well as the cross-covariances all have the forms claimed. We still need a theorem ssaying that if x_1 and x_2 are jointly gaussian and have given marginals and covariances, then the joint distribution is what we'd expect to have... $\square$