Module 9: The Multivariate Normal Distribution

Rebecca C. Steorts

Hoff, Section 7.4

Announcements

1. The last day of classes with be April 16, 2019
2. There will be a special lecture on April 18, 2019 by one of my PhD students on mixture models (abstract/title forthcoming).
3. OH will be regularly scheduled until the final exam, April 29, 2019.
4. Your lab sections will serve as extra OH by your TAs until April 29, 2019.
5. The final exam will be April 29, 2019, 9 AM – noon (Old Chem 116).

Agenda

- Moving from univariate to multivariate distributions.
- The multivariate normal (MVN) distribution.
- Conjugate for the MVN distribution.
- The inverse Wishart distribution.
- Conjugate for the MVN distribution (but on the covariance matrix).
- Combining the MVN with inverse Wishart.
- See Chapter 7 (Hoff) for a review of the standard Normal density.

Example: Reading Comprehension

A sample of 22 children are given reading comprehension tests before and after receiving a particular instructional method.[1]

Each student i will then have two scores, $Y_{i,1}$ and $Y_{i,2}$ denoting the pre- and post-instructional scores respectively.

Denote each student's pair of scores by the vector $\boldsymbol{Y}_i$

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \text{score on first test} \\ \text{score on second test} \end{pmatrix}$$

where $i = 1, \ldots, n$ and $p = 2$.

[1] This example follows Hoff (Section 7.4, p. 112).

Example: Reading Comprehension

What does this data look like that is observed?

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{n1} \\ x_{21} & x_{22} & \dots & x_{n2} \\ x_{i1} & x_{i2} & \dots & x_{ni} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

- A row of $\boldsymbol{X}_{n\times p}$ represents a covariate we might be interested in, such as age of a person.
- Denote x_i as the ith row vector of the $X_{n\times p}$ matrix.

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

where its dimension is $p \times 1$.

Example: Reading Comprehension

We may be interested in the population mean $\boldsymbol{\mu}_{p\times 1}$.

$$E[\boldsymbol{Y}] =: E[\boldsymbol{Y}_i] = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

We also may be interested in the population covariance matrix, Σ.

$$\Sigma = Cov(\boldsymbol{Y}) = \begin{pmatrix} E[Y_1^2] - E[Y_1]^2 & E[Y_1 Y_2] - E[Y_1]E[Y_2] \\ E[Y_1 Y_2] - E[Y_1]E[Y_2] & E[Y_2^2] - E[Y_2]^2 \end{pmatrix} \tag{1}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix} \tag{2}$$

Remark: $Cov(Y_1) = Var(Y_1) = \sigma_1^2$. $\qquad Cov(Y_1, Y_2) = \sigma_{1,2}$.

General Notation

Assume that $\boldsymbol{y}_{p\times 1} \sim (\mu_{p\times 1}, \Sigma_{p\times p})$.

$$\boldsymbol{y}_{p\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}.$$

$$\boldsymbol{\mu}_{p\times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

$$\Sigma_{p\times p} = Cov(\boldsymbol{y}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}.$$

Linear Algebra Background

Suppose matrix A is invertible. The

$$\det(A) = \sum_{i=1}^{j=n} a_{ij} A_{ij}.$$

I recommend using the det() commend in R.

Suppose now we have a square matrix $H_{p \times p}$.

$$\text{trace}(H) = \sum_{i} h_{ii},$$

where h_{ii} are the diagonal elements of H.

Linear Algebra Tricks

Suppose that A is $n \times n$ matrix and suppose that B is a $n \times n$ matrix.

Lemma 1:

$$tr(AB) = tr(BA)$$

Proof: Exercise.

Lemma 2:

Suppose $\boldsymbol{x}$ is a vector.

$$\boldsymbol{x}^T A\boldsymbol{x} = tr(\boldsymbol{x}^T A\boldsymbol{x}) = tr(\boldsymbol{x}\boldsymbol{x}^T A) = tr(A\boldsymbol{x}\boldsymbol{x}^T)$$

Proof: Exercise.

Why are these useful? We'll come back to this later in the module.

Notation

- MVN is generalization of univariate normal.
- For the MVN, we write $\boldsymbol{y} \sim \mathcal{MVN}(\boldsymbol{\mu}, \Sigma)$.
- The $(i, j)^{\text{th}}$ component of Σ is the covariance between Y_i and Y_j (so the diagonal of Σ gives the component variances).

Example: $Cov(Y_1, Y_2)$ is just one element of the matrix Σ.

Multivariate Normal

Just as the probability density of a scalar normal is

$$p(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, \tag{3}$$

the probability density of the multivariate normal is

$$p(\vec{x}) = (2\pi)^{-p/2} (\det\Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}. \tag{4}$$

Univariate normal is special case of the multivariate normal with a one-dimensional mean "vector" and a one-by-one variance "matrix."

Standard Multivariate Normal Distribution

Consider

$$Z_1, \ldots, Z_n \overset{iid}{\sim} N(0, 1).$$

Show that

$$Z_1, \ldots, Z_n \overset{iid}{\sim} MVN(0, I).$$

$$\begin{aligned} f_z(z) &= \prod_{i=1}^{n} (2\pi)^{-1/2} e^{-z_i^2/2} && (5) \\ &= (2\pi)^{-n/2} e^{\sum_i -z_i^2/2} && (6) \\ &= (2\pi)^{-n/2} e^{-z^T z/2}. && (7) \end{aligned}$$

Exercise: Why does $\sum_i -z_i^2 = -z^T z$?

We have just showed that $Z_1, \ldots, Z_n \overset{iid}{\sim} MVN(0, I)$.

Conjugate to MVN

Suppose that

$$X_1 \dots X_n \mid \theta \overset{iid}{\sim} MVN(\theta, \Sigma).$$

Let

$$\pi(\boldsymbol{\theta}) \sim MVN(\boldsymbol{\mu}, \Omega).$$

What is the full conditional distribution of $\boldsymbol{\theta} \mid \boldsymbol{x}, \Sigma$?

Prior

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-p/2} \det \Omega^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\} \tag{8}$$

$$\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T \Omega^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\} \tag{9}$$

$$\propto \exp -\frac{1}{2}\left\{\boldsymbol{\theta}^T \Omega^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\Omega^{-1}\mu + \mu^T\Omega^{-1}\mu\right\} \tag{10}$$

$$\propto \exp -\frac{1}{2}\left\{\boldsymbol{\theta}^T \Omega^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\Omega^{-1}\mu\right\} \tag{11}$$

$$= \exp -\frac{1}{2}\left\{\boldsymbol{\theta}^T A_o\boldsymbol{\theta} - 2\boldsymbol{\theta}^T b_o\right\} \tag{12}$$

$\pi(\boldsymbol{\theta}) \sim MVN(\boldsymbol{\mu}, \Omega)$ implies that $A_o = \Omega^{-1}$ and $b_o = \Omega^{-1}\mu$.

Likelihood

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma) = \prod_{i=1}^{n} (2\pi)^{-p/2} \det \Sigma^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \boldsymbol{\theta})^T \Sigma^{-1} (x_i - \boldsymbol{\theta})\right\} \tag{13}$$

$$\propto \exp -\frac{1}{2}\left\{\sum_i x_i^T \Sigma^{-1} x_i - 2\sum_i \boldsymbol{\theta}^T \Sigma^{-1} x_i + \sum_i \boldsymbol{\theta}^T \Sigma^{-1} \boldsymbol{\theta}\right\} \tag{14}$$

$$\propto \exp -\frac{1}{2}\left\{-2\boldsymbol{\theta}^T \Sigma^{-1} n\bar{x} + n\boldsymbol{\theta}^T \Sigma^{-1} \boldsymbol{\theta}\right\} \tag{15}$$

$$\propto \exp -\frac{1}{2}\left\{-2\boldsymbol{\theta}^T b_1 + \boldsymbol{\theta}^T A_1 \boldsymbol{\theta}\right\}, \tag{16}$$

where

$$b_1 = \Sigma^{-1} n\bar{x}, \quad A_1 = n\Sigma^{-1}$$

and

$$\bar{x} := (\frac{1}{n}\sum_i x_{i1}, \ldots, \frac{1}{n}\sum_i x_{ip})^T.$$

Full conditional

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{x}, \Sigma) &\propto p(\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma) \times p(\boldsymbol{\theta}) && (17)\\
&\propto \exp -\frac{1}{2}\Big\{-2\boldsymbol{\theta}^T b_1 + \boldsymbol{\theta}^T A_1 \boldsymbol{\theta}\Big\} && (18)\\
&\times \exp -\frac{1}{2}\Big\{\boldsymbol{\theta}^T A_o \boldsymbol{\theta} - 2\boldsymbol{\theta}^T b_o\Big\} && (19)\\
&\propto \exp\{\boldsymbol{\theta}^T b_1 - \frac{1}{2}\boldsymbol{\theta}^T A_1 \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T A_o \boldsymbol{\theta} + \boldsymbol{\theta}^T b_o\} && (20)\\
&\propto \exp\{\boldsymbol{\theta}^T (b_o + b_1) - \frac{1}{2}\theta^T (A_o + A_1)\theta\} && (21)
\end{aligned}
$$

Full conditional

From the previous slide, recall that

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \Sigma) \propto \exp\{\boldsymbol{\theta}^T (b_o + b_1) - \frac{1}{2}\theta^T (A_o + A_1)\theta\}$$

Using the kernel of the multivariate normal, we can now find the posterior mean and the posterior covariance:

Then

$$A_n = A_o + A_1 = \Omega^{-1} + n\Sigma^{-1}$$

and

$$b_n = b_o + b_1 = \Omega^{-1}\mu + \Sigma^{-1} n\bar{x}$$

$$\boldsymbol{\theta} \mid \boldsymbol{x}, \Sigma \sim MVN(A_n^{-1} b_n, A_n^{-1}) = MVN(\mu_n, \Sigma_n).$$

Interpretations

$$\boldsymbol{\theta} \mid \boldsymbol{x}, \Sigma \sim MVN(A_n^{-1} b_n, A_n^{-1}) = MVN(\mu_n, \Sigma_n)$$

$$\mu_n = A_n^{-1} b_n = [\Omega^{-1} + n\Sigma^{-1}]^{-1}(b_o + b_1) \tag{22}$$

$$= [\Omega^{-1} + n\Sigma^{-1}]^{-1}(\Omega^{-1}\mu + \Sigma^{-1} n\bar{x}) \tag{23}$$

$$\Sigma_n = A_n^{-1} = [\Omega^{-1} + n\Sigma^{-1}]^{-1} \tag{24}$$

inverse Wishart distribution

Suppose $\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1})$ where ν_o is a scalar and S_o^{-1} is a matrix.

Then

$$p(\Sigma) \propto \det(\Sigma)^{-(\nu_o+p+1)/2} \times \exp\{-\text{tr}(S_o\Sigma^{-1})/2\}$$

For the full distribution, see Hoff, Chapter 7 (p. 110).

inverse Wishart distribution

- The inverse Wishart distribution is the multivariate version of the Gamma distribution.
- The full hierarchy we're interested in is

$$\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma \sim MVN(\boldsymbol{\theta}, \Sigma).$$

$$\boldsymbol{\theta} \sim MVN(\mu, \Omega)$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

We first consider the conjugacy of the MVN and the inverse Wishart, i.e.

$$\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma \sim MVN(\boldsymbol{\theta}, \Sigma).$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}).$$

Continued

What about $p(\Sigma \mid \boldsymbol{x}, \boldsymbol{\theta}) \propto p(\Sigma) \times p(\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma)$. Let's first look at

$$
\begin{aligned}
& p(\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma) && (25) \\
& \propto \det(\Sigma)^{-n/2} \exp\{-\sum_i (\boldsymbol{x}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})/2\} && (26) \\
& \propto \det(\Sigma)^{-n/2} \exp\{-tr(\sum_i (\boldsymbol{x}_i - \boldsymbol{\theta})(\boldsymbol{x}_i - \boldsymbol{\theta})^T \Sigma^{-1}/2)\} && (27) \\
& \propto \det(\Sigma)^{-n/2} \exp\{-\mathrm{tr}(S_\theta \Sigma^{-1}/2)\} && (28)
\end{aligned}
$$

where $S_\theta = \sum_i (\boldsymbol{x}_i - \boldsymbol{\theta})(\boldsymbol{x}_i - \boldsymbol{\theta})^T$.

Note that

$$\sum_k b_k^T A b_k = tr(BB^T A),$$

where B is the matrix whose kth row is b_k. (Here we are applying Lemma 2.)

Continued

Now we can calculate $p(\Sigma \mid \boldsymbol{x}, \boldsymbol{\theta})$

$$
\begin{aligned}
& p(\Sigma \mid \boldsymbol{x}, \boldsymbol{\theta}) && (29)\\
& \quad = p(\Sigma) \times p(\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma) && (30)\\
& \quad \propto \det(\Sigma)^{-(\nu_o+p+1)/2} \times \exp\{-\mathrm{tr}(S_o\Sigma^{-1})/2\} && (31)\\
& \qquad \times \det(\Sigma)^{-n/2} \exp\{-\mathrm{tr}(S_\theta\Sigma^{-1})/2\} && (32)\\
& \quad \propto \det(\Sigma)^{-(\nu_o+n+p+1)/2} \exp\{-\mathrm{tr}((S_o + S_\theta)\Sigma^{-1})/2\} && (33)
\end{aligned}
$$

This implies that

$$\Sigma \mid, \boldsymbol{x}\boldsymbol{\theta} \sim \text{inverseWishart}(\nu_o + n, [S_o + S_\theta]^{-1} =: S_n)$$

Continued

Suppose that we wish now to take

$$\boldsymbol{\theta} \mid \boldsymbol{x}, \Sigma \sim MVN(\mu_n, \Sigma_n)$$

(which we finished an example on earlier). Now let

$$\Sigma \mid \boldsymbol{x}, \boldsymbol{\theta} \sim \text{inverseWishart}(\nu_n, S_n^{-1})$$

There is no closed form expression for this posterior. Solution?

Gibbs sampler

Suppose the Gibbs sampler is at iteration s.

1. Sample $\boldsymbol{\theta}^{(s+1)}$ from it's full conditional:
   a) Compute μ_n and Σ_n from $\boldsymbol{X}$ and $\Sigma^{(s)}$
   b) Sample $\boldsymbol{\theta}^{(s+1)} \sim MVN(\mu_n, \Sigma_n)$
2. Sample $\Sigma^{(s+1)}$ from its full conditional:
   a) Compute S_n from $\boldsymbol{x}$ and $\theta^{(s+1)}$
   b) Sample $\Sigma^{(s+1)} \sim \text{inverseWishart}(\nu_n, S_n^{-1})$