In-class exercise; introduction to multivariate normal

Dr. Olanrewaju Michael Akande

Feb 14, 2020

Announcements

- No homework this week.
- Midterm in three weeks (might seem like a lot but it is NOT!).
- Spend time practicing how to manipulate the univariate and multivariate normal distributions.

Outline

- In-class exercise
- Multivariate normal distribution

In-class exercise

- Your friend agrees to conduct a poll for you, free of charge (lucky you!).
- You give the following instructions: "Please ask about 25 people whether they are in favor of more gun control, and report back to me the number who are in favor."
- After a few days your friend returns with the poll results: there were $y = 20$ in favor. "
- You then ask, "how many people did you ask?"Your friend says, "ummm, I dunno. You didn't ask me to record that. All I know is that it was about 25."
- What model can we use to do inference here?
- **To be done on the board.**

Participation exercise

- You will work in groups of three. Work with the three students closest to you. Do the following:

  1. Using the full conditionals on the board, write a Gibbs sampler to sample from the joint posterior of N and θ, using a starting value of $N = 50$ and $\theta = 0.05$. Set burn-in to 2000 and then proceed to generate 10000 draws.

  2. Look at the posterior densities for both parameters. Describe the distributions.

  3. Give the quantile-based equal-tailed posterior credible interval for θ, rounded to two decimal places.

  4. What is the probability that exactly 20 people were polled? What can you takeaway from this?

  5. What is the probability that exactly 25 people were polled? What can you takeaway from this?

Multivariate data

- So far we have only considered basic models with scalar/univariate outcomes, $Y_1, \ldots, Y_n$.
- In practice however, outcomes of interest are actually often multivariate, e.g.,
  - Repeated measures of weight over time in a weight loss study
  - Measures of multiple disease markers
  - Tumor counts at different locations along the intestine
- Longitudinal data is just a special case of multivariate data.
- Interest then is often on how multiple outcomes are correlated, and on how that correlation may change across outcomes or time points.

MULTIVARIATE NORMAL DISTRIBUTION

- The most common model for multivariate outcomes is the multivariate normal distribution.
- Next week, we will do actual inference with the multivariate normal distribution.
- We will explore the common choices for prior distributions and then derive the corresponding posterior distributions.
- Today, we'll start slow and simply explore some properties of the multivariate normal distribution.
- Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$, where p represents the dimension of the multivariate outcome variable for a single unit of observation.
- For multiple observations, $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{ip})^T$ for $i = 1, \ldots, n$.

Multivariate normal distribution

- $\boldsymbol{Y}$ follows a multivariate normal distribution, that is, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, if

$$f(\boldsymbol{y}) = \frac{1}{\sqrt{2\pi}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right\},$$

  where $|\Sigma|$ denotes the determinant of A.

- $\boldsymbol{\mu}$ is the $p \times 1$ mean vector, that is, $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y}] = \{\mathbb{E}[Y_1], \dots, \mathbb{E}[Y_p]\} = (\mu_1, \dots, \mu_p)^T$.

- Σ is the $p \times p$ **positive semi-definite** covariance matrix, that is, $\Sigma = \{\sigma_{jk}\}$, where σ_{jk} denotes the covariance between Y_j and Y_k.

- Note that $Y_1, \dots, Y_p$ may be linearly dependent depending on the structure of Σ, which characterizes association between them.

- For each $j = 1, \dots, p$, $Y_j \sim \mathcal{N}(\mu_j, \sigma_{jj})$.

Bivariate normal distribution

- In the bivariate case, we have

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left[\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} = \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22} = \sigma_2^2 \end{pmatrix}\right],$$

  and $\sigma_{12} = \sigma_{21} = \mathbb{C}\mathrm{ov}[Y_1, Y_2]$.

- The correlation between Y_1 and Y_2 is defined as

$$\rho_{1,2} = \frac{\mathbb{C}\mathrm{ov}[Y_1, Y_2]}{\sqrt{\mathbb{V}\mathrm{ar}[Y_1]}\sqrt{\mathbb{V}\mathrm{ar}[Y_2]}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}.$$

- $-1 \leq \rho_{1,2} \leq 1$.

- Correlation coefficient is free of the measurement units.

Back to the multivariate normal

- There are many special properties of the multivariate normal as we will see as we continue to work with the distribution.

- First, dependence between any Y_j and Y_k does not depend on the other $p-2$ variables.

- Second, while generally, **independence implies zero covariance**, for the normal family, the converse is also true. That is, **independence implies zero covariance**.

- Thus, the covariance Σ carries a lot of information about marginal relationships, especially **marginal independence**.

- If $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$, that is, $\epsilon_1, \ldots, \epsilon_p \overset{iid}{\sim} \mathcal{N}(0, 1)$, then

$$\boldsymbol{Y} = \boldsymbol{\mu} + A\boldsymbol{\epsilon} \Rightarrow \ \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

  holds for any matrix square root A of Σ, that is, $AA^T = \Sigma$ (see Cholesky decomposition).

Conditional distributions

- Partition $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$ as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right],$$

  where

  - $\boldsymbol{Y}_1$ and $\boldsymbol{\mu}_1$ are $q \times 1$, and $\boldsymbol{Y}_2$ and $\boldsymbol{\mu}_2$ are $(p-q) \times 1$;
  - Σ_{11} is $q \times q$, and Σ_{22} is $(p-q) \times (p-q)$, with $\Sigma_{22} > 0$.

- Then, it turns out that

$$\boldsymbol{Y}_1 | \boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

- That is, the conditional distribution of $\boldsymbol{Y}_1$ given $\boldsymbol{Y}_2$ is also normal!

- Marginal distributions are once again normal, that is,

$$\boldsymbol{Y}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_1, \Sigma_{11}\right); \quad \boldsymbol{Y}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_2, \Sigma_{22}\right).$$

Conditional distributions

- In the bivariate normal case with

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left[\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} = \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22} = \sigma_2^2 \end{pmatrix}\right],$$

we have

$$Y_1 | Y_2 = y_2 \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2}(y_2 - \mu_2), \sigma_1 - \frac{\sigma_{12}^2}{\sigma_2}\right).$$

which can also be written as

$$Y_1 | Y_2 = y_2 \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(y_2 - \mu_2), (1 - \rho^2)\sigma_1^2\right).$$

Working with normal distributions

- Three real (univariate) random quantities x, y and z have a joint normal distribution given by $p(x, y, z) = p(y|x)p(x|z)p(z)$.
- Suppose
  - $p(y|x) = \mathcal{N}(x, w)$ independently of z, for some known variance w;
  - $p(x|z) = \mathcal{N}(\theta z, v)$ for some known parameter θ, and known variance v; and
  - $p(z) = \mathcal{N}(m, M)$, with some known mean m, and known variance M.
- What is
  - $p(x)$? $p(y)$?
  - $p(x|y)$? $p(z|x)$?
- **To be done on the board.**