The multivariate normal distribution

Patrick Breheny

September 6

Introduction

- Today we will introduce the multivariate normal distribution and attempt to discuss its properties in a fairly thorough manner
- The multivariate normal distribution is by far the most important multivariate distribution in statistics
- It's important for all the reasons that the one-dimensional Gaussian distribution is important, but even more so in higher dimensions because many distributions that are useful in one dimension do not easily extend to the multivariate case

Motivation

- In the univariate case, the family of normal distributions can be constructed from the standard normal distribution through the location-scale transformation $\mu + \sigma Z$, where $Z \sim \mathrm{N}(0, 1)$; the resulting random variable has a $\mathrm{N}(\mu, \sigma^2)$ distribution
- A similar approach can be taken with the multivariate normal distribution, although some care needs to be taken with regard to whether the resulting variance is singular or not

Standard normal

- First, the easy case: if $Z_1, \ldots, Z_r$ are mutually independent and each follows a standard normal distribution, the random vector $\mathbf{z}$ is said to follow an r-variate standard normal distribution, denoted $\mathbf{z} \sim \mathrm{N}_r(\mathbf{0}, \mathbf{I}_r)$
- Remark: For multivariate normal distributions and identity matrices, I will usually leave off the subscript from now on when it is either unimportant or able to be figured out from context
- If $\mathbf{z} \sim \mathrm{N}_r(\mathbf{0}, \mathbf{I})$, its density is

$$p(\mathbf{z}) = (2\pi)^{-r/2} \exp\{-\tfrac{1}{2}\mathbf{z}^\top \mathbf{z}\}$$

Multivariate possibilities

- Like the univariate case, we can construct multivariate distributions through linear combinations
- Before we define the multivariate normal distribution, however, note that there is no guarantee that the dimension remains the same in such a transformation:
    - Suppose $z_1, z_2, z_3 \stackrel{\perp\!\!\!\perp}{\sim} \mathrm{N}(0,1)$
    - The dimension could decrease: $x_1 = z_1 + 2z_3, x_2 = -z_2$
    - Or increase:

$$
\begin{aligned}
x_1 &= z_1 + 2z_2 \\
x_2 &= z_1 - z_2 \\
x_3 &= z_2 - z_3 \\
x_4 &= z_1 + z_2 + z_3
\end{aligned}
$$

Multivariate normal distribution

- **Definition:** Let $\mathbf{x}$ be a $d \times 1$ random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, where $\mathrm{rank}(\boldsymbol{\Sigma}) = r > 0$. Let $\boldsymbol{\Gamma}$ be a $r \times d$ matrix such that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^\top \boldsymbol{\Gamma}$. Then $\mathbf{x}$ is said to have a *d-variate normal distribution of rank r* if its distribution is the same as that of the random vector $\boldsymbol{\mu} + \boldsymbol{\Gamma}^\top \mathbf{z}$, where $\mathbf{z} \sim \mathrm{N}_r(\mathbf{0}, \mathbf{I})$.
- This is typically denoted $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Density

- Suppose $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that $\boldsymbol{\Sigma}$ is full rank; then $\mathbf{x}$ has a density:

  $$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\},$$

  where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$
- We will not really concern ourselves with determinants and their properties in this course, although it is worth pointing out that if $\boldsymbol{\Sigma}$ is singular, then $|\boldsymbol{\Sigma}| = 0$ and the above result does not hold (or even make sense)

Singular case

- In fact, if $\mathbf{\Sigma}$ is singular, then $\mathbf{x}$ does not even *have* a density
- This is connected to our earlier discussion of the Lebesgue decomposition theorem: if $\mathbf{\Sigma}$ is singular, then the distribution of $\mathbf{x}$ has a singular component (i.e., $\mathbf{x}$ is not absolutely continuous)
- This is the reason why the definition of the MVN might seem somewhat roundabout – we can't just say that the random variable has a certain density, but must instead say that it has the same distribution as $\boldsymbol{\mu} + \mathbf{\Gamma}^\top \mathbf{z}$, where $\mathbf{z}$ has a well-defined density

Moment generating function

- For this reason, when working with multivariate normal distributions or showing that some random variable $\mathbf{y}$ follows a multivariate normal distribution, it is often easier to work with moment generating functions or characteristic functions, which are well-defined even if $\mathbf{\Sigma}$ is singular
- If $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \mathbf{\Sigma})$, then its moment generating function is

$$m(\mathbf{t}) = \exp\{\mathbf{t}^\top \boldsymbol{\mu} + \tfrac{1}{2}\mathbf{t}^\top \mathbf{\Sigma} \mathbf{t}\},$$

  where $\mathbf{t} \in \mathbb{R}^d$
- We'll come back to its characteristic function in a future lecture

Partitioned matrices

- The concept of partitioning a matrix will come up often
- The idea of a partitioned matrix is to think of a large matrix as a collection of smaller submatrices:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 2 & 7 \\ 1 & 5 & 6 & 2 \\ 3 & 3 & 4 & 5 \\ 3 & 3 & 6 & 7 \end{bmatrix}$$

can be partitioned into four 2×2 blocks

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \text{ where } \mathbf{A}_{11} = \begin{bmatrix} 1 & 2 \\ 1 & 5 \end{bmatrix}, \mathbf{A}_{12} = \begin{bmatrix} 2 & 7 \\ 6 & 2 \end{bmatrix}, \ldots$$

Transposing partitioned matrices

- The transpose of a partitioned matrix is

$$\mathbf{A}^\top = \begin{bmatrix} \mathbf{A}_{11}^\top & \mathbf{A}_{21}^\top \\ \mathbf{A}_{12}^\top & \mathbf{A}_{22}^\top \end{bmatrix}$$

- Note that if $\mathbf{A}$ is symmetric, as in the case of a covariance matrix or matrix of second derivatives, then

$$\mathbf{A}_{12}^\top = \mathbf{A}_{21}$$

Independence

- Before moving on, let us note that there is a connection between covariance and independence in the multivariate normal distribution
- **Theorem:** Suppose $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\mathbf{x} = [\mathbf{x}_1 \, \mathbf{x}_2]^\top$ and the corresponding off-diagonal of $\boldsymbol{\Sigma}_{12}$ is zero, then $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent.
- In particular, if $\boldsymbol{\Sigma}$ is a diagonal matrix, then $x_1, \ldots, x_n$ are mutually independent

Independence (caution)

- It is worth pointing out a common mistake here: $\text{Cov}(X_1, X_2) = 0 \implies X_1 \perp\!\!\!\perp X_2$ only if X_1 and X_2 are *multivariate normal*
- For example, suppose $X \sim N(0, 1)$ and $Y = \pm X$, each with probability $\frac{1}{2}$
- X and Y are both normally distributed, and $\text{Cov}(X, Y) = 0$, but they are clearly not independent

Main result

- A very important property of the multivariate normal distribution is that its linear combinations are also normally distributed
- **Theorem:** Let $\mathbf{b}$ be a $k \times 1$ vector of constants, $\mathbf{B}$ a $k \times d$ matrix of constants, and $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$\mathbf{b} + \mathbf{B}\mathbf{x} \sim \mathrm{N}_k(\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top).$$

Corollary

- A useful corollary of this result is that we can always "standardize" a variable with an MVN distribution
- Let's consider the full-rank case first (i.e., $\boldsymbol{\Sigma}$ is nonsingular and positive definite, and so is $\boldsymbol{\Sigma}^{-1}$)
- **Corollary:** Let $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim \mathrm{N}_d(\mathbf{0}, \mathbf{I}),$$

where $\boldsymbol{\Sigma}^{-1/2}$ is the square root of $\boldsymbol{\Sigma}^{-1}$.

Corollary: Low rank case

- If $\boldsymbol{\Sigma}$ is singular, then $\boldsymbol{\Sigma}^{-1/2}$ does not exist, although we can still standardize the distribution
- **Corollary:** Let $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is rank r with $\boldsymbol{\Gamma}^\top \boldsymbol{\Gamma} = \boldsymbol{\Sigma}$. Then

$$(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top)^{-1}\boldsymbol{\Gamma}(\mathbf{x} - \boldsymbol{\mu}) \sim \mathrm{N}_r(\mathbf{0}, \mathbf{I}).$$

Main result

- In the univariate case, if $Z \sim \mathrm{N}(0,1)$, then Z^2 follows a distribution known as the χ^2 distribution
- Furthermore, if $Z_1, \ldots, Z_n$ are mutually independent and each $Z_i \sim \mathrm{N}(0,1)$, then $\sum_i Z_i^2 \sim \chi^2_n$, where χ^2_n denotes the χ^2 distribution with n degrees of freedom
- Thus, it is a straightforward consequence of our previous corollaries that if $\mathbf{x} \sim \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is nonsingular,

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} \sim \chi^2_d$$

Main result (low rank)

- Similarly, it is always the case that if $\mathbf{x} \sim \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^\top \boldsymbol{\Gamma}$, then

$$\mathbf{x}^\top \boldsymbol{\Sigma}^- \mathbf{x} \sim \chi^2_r,$$

  where r is the rank of $\boldsymbol{\Sigma}$ and

$$\boldsymbol{\Sigma}^- = \boldsymbol{\Gamma}^\top (\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top)^{-1} (\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top)^{-1} \boldsymbol{\Gamma}$$

- As discussed in our review last time, $\boldsymbol{\Sigma}^-$ is a quantity known as a *generalized inverse*, which you'll learn more about in the linear models course

Non-central chi square distribution

- If $\boldsymbol{\mu} \neq \mathbf{0}$, then the quadratic form follows something called a non-central χ^2 distribution
- If $Z_1, \ldots, Z_n \overset{\perp\!\!\!\perp}{\sim} N(\mu_i, 1)$, then the distribution of $\sum_i Z_i^2$ is known as the noncentral χ_n^2 distribution with noncentrality parameter $\sum_i \mu_i^2$
- Thus, if $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} \sim \chi_d^2(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}),$$

or

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-} \mathbf{x} \sim \chi_r^2(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-} \boldsymbol{\mu})$$

if $\boldsymbol{\Sigma}$ is singular

Marginal distributions

- Finally, let us consider some results related to partitions of the multivariate normal distribution:

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

- Conveniently, the marginal distributions are exactly what you would intuitively think they should be:

$$\mathbf{x}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

Conditional

- A more complicated question: what is the distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$?
- This gets messy if $\boldsymbol{\Sigma}$ is singular, but if $\boldsymbol{\Sigma}$ is full rank, then

$$\mathbf{x}_1 | \mathbf{x}_2 \sim \mathrm{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

- As mentioned earlier, note that if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, then $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent and $\mathbf{x}_1 | \mathbf{x}_2 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$;

Schur complement

- The quantity $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ is known in linear algebra as the *Schur complement*; it comes up all the time in statistics and we will see it repeatedly in this course
- It is the **inverse** of the $(1, 1)$ block of $\boldsymbol{\Sigma}^{-1}$; more explicitly, letting $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$,

$$\boldsymbol{\Theta}_{11}^{-1} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

- Conceptually, it represents the reduction in the variability of $\mathbf{x}_1$ that we achieve by learning $\mathbf{x}_2$ (or equivalently, the increase in our uncertainty about $\mathbf{x}_1$ if we don't know $\mathbf{x}_2$)

Precision matrix

- The inverse of the covariance matrix, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, is known as the *precision matrix* and is a rather interesting quantity in its own right
- In fact, many statistical procedures are more concerned with estimating $\mathbf{\Theta}$ than $\mathbf{\Sigma}$
- One key reason for this is that $\mathbf{\Theta}$ encodes conditional independence relationships that are often of interest in learning the structure of $\mathbf{x}$ in terms of which how variables are related to each other

Conditional independence result

- Suppose we partition $\mathbf{x}$ into $\mathbf{x}_1$, containing two variables of interest, and $\mathbf{x}_2$ containing the remaining variables
- Then by the results we've obtained already, if $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{x}_1 | \mathbf{x}_2$ is multivariate normal with covariance matrix $\boldsymbol{\Theta}_{11}^{-1}$
- Thus, if any off-diagonal element of $\boldsymbol{\Theta}$ is zero, then the corresponding variables are conditionally independent given the remaining variables
- This is of interest in many statistical problems

Example: $X \to Y \to Z$

```
n <- 100000
x <- rnorm(n)
y <- x + rnorm(n)
z <- y + rnorm(n)
cor(cbind(x, y, z))
#           x         y         z
# x 1.0000000 0.7065722 0.5763312
# y 0.7065722 1.0000000 0.8173433
# z 0.5763312 0.8173433 1.0000000
cor(cbind(x, y, z)) |> solve()
#              x         y            z
# x  1.996998254 -1.416830  0.007104064
# y -1.416829854  4.017715 -2.467289446
# z  0.007104064 -2.467289  3.012528287
```

Application

- One application of this idea is in learning gene regulatory networks
- Suppose the expression levels of various genes follow a multivariate normal distribution (at least approximately)
- Learning which elements of $\mathbf{\Theta}$ are nonzero corresponds to learning which pairs of genes have a direct relationship with one another, as opposed to being merely correlated through the effects of other genes that affect them both