The multivariate normal distribution

Patrick Breheny

September 2

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

Inverse

- Before we get to the multivariate normal distribution, let's review some important results from linear algebra that we will use throughout the course, starting with inverses
- **Definition:** The *inverse* of an $n \times n$ matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix satisfying $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.
- Note: We're sort of getting ahead of ourselves by saying that $\mathbf{A}^{-1}$ is "the" matrix satisfying $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$, but it is indeed the case that if a matrix has an inverse, the inverse is unique

Singular matrices

- However, not all matrices have inverses; for example

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

- There does not exist a matrix such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_2$
- Such matrices are said to be *singular*
- Remark: Only square matrices have inverses; an $n \times m$ matrix $\mathbf{A}$ might, however, have a *left inverse* (satisfying $\mathbf{B}\mathbf{A} = \mathbf{I}_m$) or *right inverse* (satisfying $\mathbf{A}\mathbf{B} = \mathbf{I}_n$)

Positive definite

- A related notion is that of a "positive definite" matrix, which applies to symmetric matrices
- **Definition:** A symmetric $n \times n$ matrix $\mathbf{A}$ is said to be *positive definite* if for all $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \qquad \text{if } \mathbf{x} \neq 0$$

- The two notions are related in the sense that if $\mathbf{A}$ is positive definite, then (a) $\mathbf{A}$ is not singular and (b) $\mathbf{A}^{-1}$ is also positive definite
- If $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$, then $\mathbf{A}$ is said to be *positive semidefinite*
- In statistics, these classifications are particularly important for variance-covariance matrices, which are always positive semidefinite (and positive definite, if they aren't singular)

Square root of a matrix

- These concepts are important with respect to knowing whether a matrix has a "square root"
- **Definition:** An $n \times n$ matrix $\mathbf{A}$ is said to have a *square root* if there exists a matrix $\mathbf{B}$ such that $\mathbf{BB} = \mathbf{A}$.
- **Theorem:** Let $\mathbf{A}$ be a positive definite matrix. Then there exists a unique matrix $\mathbf{A}^{1/2}$ such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$.
- Positive semidefinite matrices have square roots as well, although they aren't necessarily unique

Rank

- One additional linear algebra concept we need is the rank of a matrix (there are many ways of defining rank; all are equivalent)
- **Definition:** The *rank* of a matrix is the dimension of its largest nonsingular submatrix.
- For example, the following 3×3 matrix is singular, but contains a nonsingular 2×2 submatrix, so its rank is 2:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & \not{3} \\ \not{2} & \not{4} & \not{6} \\ 1 & 0 & \not{1} \end{bmatrix}$$

- Note that a nonsingular $n \times n$ matrix has rank n, and is said to be *full rank*

Rank and multiplication

- There are many results and theorems involving rank; we're not going to cover them all, but it is important to know that rank cannot be increased through the process of multiplication
- **Theorem:** For any matrices $\mathbf{A}$ and $\mathbf{B}$ with appropriate dimensions, $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{A})$ and $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{B})$.
- In particular, $\text{rank}(\mathbf{A}^\top\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}^\top) = \text{rank}(\mathbf{A})$.

Expectation and variance

- Finally, we need some results on expected values of vectors and functions of vectors
- First of all, we need to define expectation and variance as they pertain to random vectors
- **Definition:** Let $\mathbf{x} = (X_1 \;\; X_2 \;\; \cdots X_d)$ denote a vector of random variables, then $\mathbb{E}(\mathbf{x}) = (\mathbb{E}X_1 \; \mathbb{E}X_2 \;\; \cdots \mathbb{E}X_d)$. Meanwhile, $\mathbb{V}\mathbf{x}$ is a $d \times d$ matrix:

$$\mathbb{V}\mathbf{x} = \mathbb{E}\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\} \text{ with elements}$$
$$(\mathbb{V}\mathbf{x})_{ij} = \mathbb{E}\left\{(X_i - \mu_i)(X_j - \mu_j)\right\},$$

where $\mu_i = \mathbb{E}X_i$. The matrix $\mathbb{V}\mathbf{x}$ is referred to as the *variance-covariance matrix* of $\mathbf{x}$.

Linear and quadratic forms

- Letting $\mathbf{A}$ denote a matrix of constants and $\mathbf{x}$ a random vector with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$,

$$\begin{aligned} \mathbb{E}(\mathbf{A}^T\mathbf{x}) &= \mathbf{A}^T\boldsymbol{\mu} \\ \mathbb{V}(\mathbf{A}^T\mathbf{x}) &= \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A} \\ \mathbb{E}(\mathbf{x}^T\mathbf{A}\mathbf{x}) &= \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}), \end{aligned}$$

where $\mathrm{tr}(\mathbf{A}) = \sum_i A_{ii}$ is the trace of $\mathbf{A}$

- Some useful facts about traces:

$$\begin{aligned} \mathrm{tr}(\mathbf{AB}) &= \mathrm{tr}(\mathbf{BA}) \\ \mathrm{tr}(\mathbf{A}+\mathbf{B}) &= \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) \\ \mathrm{tr}(c\mathbf{A}) &= c\,\mathrm{tr}(\mathbf{A}) \\ \mathrm{tr}(\mathbf{A}) &= \mathrm{rank}(\mathbf{A}) \quad \text{if } \mathbf{AA} = \mathbf{A} \end{aligned}$$

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 normal distribution

- **Definition:** Let $\mathbf{x}$ be a $d \times 1$ random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, where $\text{rank}(\boldsymbol{\Sigma}) = r > 0$. 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 $\mathbf{z} \sim \text{N}_r(\mathbf{0}, \mathbf{I})$.
- This is typically denoted $\mathbf{x} \sim \text{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

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 $\boldsymbol{\Sigma}$ is singular
- If $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then its moment generating function is

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

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

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 $\mathbf{x} = [\mathbf{x}_1 \, \mathbf{x}_2]$ 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 $\mathbf{\Sigma}$ is singular, then $\mathbf{\Sigma}^{-1/2}$ does not exist, although we can still standardize the distribution
- **Corollary:** Let $\mathbf{x} \sim \mathrm{N}_d(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is rank r with $\mathbf{\Gamma}^\top \mathbf{\Gamma} = \mathbf{\Sigma}$. Then

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

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 $\sum_i Z_i^2 \sim \chi_n^2$, where χ_n^2 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,

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

Main result (low rank)

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

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

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

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

- Here, $\mathbf{\Sigma}^-$ is a quantity known as a *generalized inverse*
- We won't discuss them any further in this course, but you can learn more about them 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 $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ 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 $\mathbf{\Sigma}$; more explicitly, letting $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$,

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

- 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, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, 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 $\boldsymbol{\Theta}$ than $\boldsymbol{\Sigma}$
- One key reason for this is that $\boldsymbol{\Theta}$ 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

- For example, suppose $X \to Y \to Z$; we could simulate this with, for example,

```
x <- rnorm(n)
y <- x + rnorm(n)
z <- y + rnorm(n)
```

- Note that $\hat{\boldsymbol{\Sigma}}_{xz}$ is not close to zero at all; X and Z are not independent and are, in fact, rather highly correlated
- However, $\hat{\boldsymbol{\Theta}}_{xz} \approx 0$; X and Z are *conditionally independent* given Y

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