Linear & Generalized Linear Models/ Advanced Regression for Independent Data

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 1: Random vectors and matrices; multivariate normal theory

Version: 09/02/2026

Table of Contents

1. Random vectors/matrices and their properties
2. The multivariate normal distribution
3. Other important distributions

Random vectors/matrices and their properties

Random vectors:

- A random vector is, as implied by the name, a vector of random variables:

$$\mathbf{x} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}.$$

- It has expectation given by:

$$\mathrm{E}[\mathbf{x}] = \begin{bmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_N] \end{bmatrix}.$$

Random matrices:

- A random matrix is, as implied by the name, a matrix of random variables:

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1K} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{NK} \end{bmatrix}.$$

- It has expectation given by:

$$\mathrm{E}[\mathbf{X}] = \begin{bmatrix} \mathrm{E}[X_{11}] & \cdots & \mathrm{E}[X_{1K}] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[X_{N1}] & \cdots & \mathrm{E}[X_{NK}] \end{bmatrix}.$$

Random vectors/matrices and their properties

Rules for expectations:

- Let $\mathbf{A}$ and $\mathbf{B}$ denote constant matrices.
- Let $\mathbf{X}$ and $\mathbf{Y}$ denote random matrices.
- Throughout this course, any presentation of matrix operations imply that the matrices must be conformable for that operation (i.e., the operation is well defined for matrices having the respective dimensions).
- The following properties hold:
  1. $\mathrm{E}[\mathbf{A}] = \mathbf{A}$.
  2. $\mathrm{E}[\mathbf{AX}] = \mathbf{A}\mathrm{E}[\mathbf{X}]$.
  3. $\mathrm{E}[\mathbf{X} + \mathbf{Y}] = \mathrm{E}[\mathbf{X}] + \mathrm{E}[\mathbf{Y}]$.
  4. $\mathrm{E}[\mathbf{AXB}] = \mathbf{A}\mathrm{E}[\mathbf{X}]\mathbf{B}$.
  5. $\mathrm{E}[\mathbf{X} + \mathbf{A}] = \mathrm{E}[\mathbf{X}] + \mathbf{A}$.
  6. $\mathrm{E}_{\mathbf{X},\mathbf{Y}}[g(\mathbf{X}, \mathbf{Y})] = \mathrm{E}_{\mathbf{X}}[\mathrm{E}_{\mathbf{Y}|\mathbf{X}}[g(\mathbf{X}, \mathbf{Y})]]$. (iterated expectation rule)
  7. $\mathrm{E}[\mathbf{Y}] = \mathrm{E}_{\mathbf{X}}[\mathrm{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}]]$. (law of total expectation)

Random vectors/matrices and their properties

Covariance:

- If $\mathbf{x}$ is a random vector, then its covariance matrix is defined as:

$$\text{Cov}[\mathbf{x}] = \begin{bmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \cdots & \text{Cov}[X_1, X_N] \\ \text{Cov}[X_2, X_1] & \text{Var}[X_2] & \cdots & \text{Cov}[X_2, X_N] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_N, X_1] & \text{Cov}[X_N, X_2] & \cdots & \text{Var}[X_N] \end{bmatrix}.$$

- I may sometimes refer to this as $\text{Var}[\mathbf{x}]$, though you should be able to tell by context that this refers to the whole variance-covariance matrix.
- Note: $\text{Cov}[\mathbf{x}] = \text{E}[(\mathbf{x} - \text{E}[\mathbf{x}])(\mathbf{x} - \text{E}[\mathbf{x}])^\intercal]$.
- Note: $\text{Cov}[\mathbf{x}] = \text{E}[\mathbf{x}\mathbf{x}^\intercal] - \text{E}[\mathbf{x}](\text{E}[\mathbf{x}])^\intercal$.

Covariance: Properties

- If $X_1, \ldots, X_N$ are independent, $\text{Cov}[\mathbf{x}]$ is diagonal.
- If $X_1, \ldots, X_N$ are independent with common variance, $\text{Cov}[\mathbf{x}] = \sigma^2 \mathbf{I}$.
- $\text{Cov}[\mathbf{x}]$ is symmetric.
- $\text{Cov}[\mathbf{x} + \mathbf{a}] = \text{Cov}[\mathbf{x}]$ for a constant vector $\mathbf{a}$.
- $\text{Cov}[\mathbf{A}\mathbf{x}] = \mathbf{A}\text{Cov}[\mathbf{x}]\mathbf{A}^\intercal$ for a constant matrix $\mathbf{A}$.
- $\text{Cov}[\mathbf{y}] = \text{E}_{\mathbf{x}}[\text{Cov}_{\mathbf{y}|\mathbf{x}}[\mathbf{y}]] + \text{Cov}_{\mathbf{x}}[\text{E}_{\mathbf{y}|\mathbf{x}}[\mathbf{y}]]$. (law of total variance)
- Since $\text{Cov}[\mathbf{x}]$ is positive semi-definite, $\text{Var}[\mathbf{a}^\intercal \mathbf{x}] = \mathbf{a}^\intercal \text{Cov}[\mathbf{x}]\mathbf{a} \geq 0$.
  - If no linear combination of the elements of $\mathbf{x}$ is constant, $\text{Cov}[\mathbf{x}]$ is in fact positive definite. For example, $\text{Cov}[(X_1, 2 - X_1)^\intercal]$ is positive semi-definite but not positive definite.

Covariance of two random vectors:

- Let $\mathbf{x}$ and $\mathbf{y}$ denote random vectors of lengths N and K, respectively. Then, $\text{Cov}[\mathbf{x}, \mathbf{y}]$ is an $N \times K$ matrix:

$$\text{Cov}[\mathbf{x}, \mathbf{y}] = \begin{bmatrix} \text{Cov}[X_1, Y_1] & \text{Cov}[X_1, Y_2] & \cdots & \text{Cov}[X_1, Y_K] \\ \text{Cov}[X_2, Y_1] & \text{Cov}[X_2, Y_2] & \cdots & \text{Cov}[X_2, Y_K] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[X_N, Y_1] & \text{Cov}[X_N, Y_2] & \cdots & \text{Cov}[X_N, Y_K] \end{bmatrix}$$

Covariance of two random vectors: Properties

- Indeed, $\text{Cov}[\mathbf{x}, \mathbf{y}] = \text{E}[(\mathbf{x} - \text{E}[\mathbf{x}])(\mathbf{y} - \text{E}[\mathbf{y}])^\intercal]$.
- As you might expect, $\text{Cov}[\mathbf{x}, \mathbf{x}] = \text{Cov}[\mathbf{x}]$.
- If $\mathbf{A}$ and $\mathbf{B}$ are constant matrices, then $\text{Cov}[\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}] = \mathbf{A}\text{Cov}[\mathbf{x}, \mathbf{y}]\mathbf{B}^\intercal$.
- Covariance matrices can be partitioned. If

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

then,

$$\text{Cov}[\mathbf{z}] = \begin{bmatrix} \text{Cov}[\mathbf{x}] & \text{Cov}[\mathbf{x}, \mathbf{y}] \\ \text{Cov}[\mathbf{y}, \mathbf{x}] & \text{Cov}[\mathbf{y}] \end{bmatrix}.$$

- $\text{Cov}[\mathbf{x}, \mathbf{y}] = (\text{Cov}[\mathbf{y}, \mathbf{x}])^\intercal$.

Theorem 1.1: Expectations of quadratic forms

Let $\mathbf{x}$ denote a random length-N vector (let $\boldsymbol{\mu} = \mathsf{E}[\mathbf{x}]$ denote the mean vector and let $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{x}]$ denote the covariance matrix). If $\mathbf{A}$ is a constant $N \times N$ matrix, then:

$$\mathsf{E}[\mathbf{x}^\intercal \mathbf{A} \mathbf{x}] = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\intercal \mathbf{A} \boldsymbol{\mu}.$$

Quick question: What are the dimensions of $\mathbf{x}^\intercal \mathbf{A} \mathbf{x}$?

Theorem 1.1: Proof

- Let $\mathbf{x}$ denote a random length-N vector (let $\boldsymbol{\mu} = \mathrm{E}[\mathbf{x}]$ and $\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{x}]$). If $\mathbf{A}$ is a constant $N \times N$ matrix, then:

$$\begin{aligned}
\mathrm{E}[\mathbf{x}^\top \mathbf{A}\mathbf{x}] &= \mathrm{E}[\mathrm{tr}(\mathbf{x}^\top \mathbf{A}\mathbf{x})] \\
&= \mathrm{E}[\mathrm{tr}(\mathbf{A}\mathbf{x}\mathbf{x}^\top)] \\
&= \mathrm{tr}(\mathrm{E}[\mathbf{A}\mathbf{x}\mathbf{x}^\top]) \\
&= \mathrm{tr}(\mathbf{A}\mathrm{E}[\mathbf{x}\mathbf{x}^\top]) \\
&= \mathrm{tr}(\mathbf{A}(\mathrm{Cov}[\mathbf{x}] + \mathrm{E}[\mathbf{x}](\mathrm{E}[\mathbf{x}])^\top)) \\
&= \mathrm{tr}(\mathbf{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top)) \\
&= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma} + \mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^\top) \\
&= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \mathrm{tr}(\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^\top) \\
&= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \mathrm{tr}(\boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}) \\
&= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}.
\end{aligned}$$

- I want you to memorize the formula (but not the proof).

Example 1.1: Unbiasedness of the sample variance

- Let $X_1, \ldots, X_N \sim (\mu, \sigma^2)$ be independent random variables. Then, we can verify that the sample variance is unbiased for σ^2:

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2\right] = \sigma^2.$$

- To see this, let $\mathbf{x} = (X_1, \ldots, X_N)^\mathsf{T}$, with $\mathrm{E}[\mathbf{x}] = \mu\mathbf{1}$ and $\mathrm{Cov}[\mathbf{x}] = \sigma^2\mathbf{I}$. Defining $\mathbf{A} = \mathbf{I} - \mathbf{J}_N/N$ (serves to subtract mean from each element of the input vector),

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \sum_{i=1}^{N}(X_i - \overline{X})^2 = (N-1)S^2.$$

- Now,

$$\mathrm{E}[(N-1)S^2] = \mathrm{E}[\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}] = \sigma^2\mathrm{tr}(\mathbf{A}) + \mu^2\mathbf{1}^\mathsf{T}\mathbf{A}\mathbf{1} = (N-1)\sigma^2.$$

Table of Contents

1 Random vectors/matrices and their properties

2 The multivariate normal distribution

3 Other important distributions

Goal:

- The multivariate normal distribution comes up a *lot* in statistics.
- Our task now is to cycle through everything you (n)ever wanted to know about it—and then some.
  - Definitions.
  - Properties.
  - Marginal and conditional distributions.
  - Quadratic forms and the χ^2 distribution.
  - The non-central χ^2 distribution.
  - The F-distribution and the non-central F-distribution.
- I will also note that many of the results on multivariate normal theory will not be invoked until we talk about hypothesis testing.

Definition 1:

- For a length-N vector $\boldsymbol{\mu}$ and $N \times N$ positive definite matrix $\boldsymbol{\Sigma}$, the random vector $\mathbf{y}$ follows a multivariate normal distribution, $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if the density function for $\mathbf{y}$ is given by:

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-N/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right).$$

Aside: Can you see why this quantity is a scalar?

The multivariate normal distribution

Definition 2:

- For a length-N vector $\boldsymbol{\mu}$ and $N \times N$ positive semi-definite matrix $\boldsymbol{\Sigma}$, the random vector $\mathbf{y}$ follows a multivariate normal distribution, $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if it has moment generating function (MGF):

$$M_{\mathbf{y}}(\mathbf{t}) = \mathrm{E}[\exp(\mathbf{t}^{\mathsf{T}}\mathbf{y})] = \exp\left(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{t}\right).$$

Definition 3:

- For a length-N vector $\boldsymbol{\mu}$ and a rank-R positive semi-definite $N \times N$ matrix $\boldsymbol{\Sigma}$, the random vector $\mathbf{y}$ follows a multivariate normal distribution, $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if for $\mathbf{z} = (Z_1, \ldots, Z_K)$ based on K independent standard normal random variables, $\mathbf{y}$ has the same distribution as $\mathbf{Az} + \boldsymbol{\mu}$, where $\mathbf{A}$ is an $N \times K$ matrix such that $\mathbf{A}\mathbf{A}^\intercal = \boldsymbol{\Sigma}$ (in fact, every valid $\boldsymbol{\Sigma}$ can be expressed as $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\intercal$ for some $\mathbf{A}$. Why?).
    - We need $K \geq R$. Truly, choosing $K = R$ gets the job done just fine. The intuition is that there are R independent sources of randomness; $\mathbf{A}$ serves to (linearly) combine them to produce the correlation structure reflected by $\boldsymbol{\Sigma}$.

Definitions:

- Definitions 1, 2, and 3 are equivalent for $\mathbf{\Sigma} \succ 0$ (recall that this is short-hand notation for positive definite).
- Definitions 2 and 3 are equivalent for $\mathbf{\Sigma} \succcurlyeq 0$ (recall that this is short-hand notation for positive semi-definite).
- Life is short; let's not prove the various equivalences (though you can look it up). The bottom line is that these various definitions are helpful at different times in different ways.

Properties: Suppose $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- $\mathrm{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\mathrm{Cov}[\mathbf{y}] = \boldsymbol{\Sigma}$. These statements are not "obvious." The choice to label the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ make it seem as if they are when in fact they are so chosen and labeled *because* of the properties.
- Odd-ordered central moments of a multivariate normal distribution are zero. This includes quantities of the form $\mathrm{E}[(X_1 - \mu_1)^3]$, $\mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)(X_3 - \mu_3)]$, and $\mathrm{E}[(X_1 - \mu_1)^2(X_2 - \mu_2)^5]$.
  - This is true, more generally, of symmetric distributions.

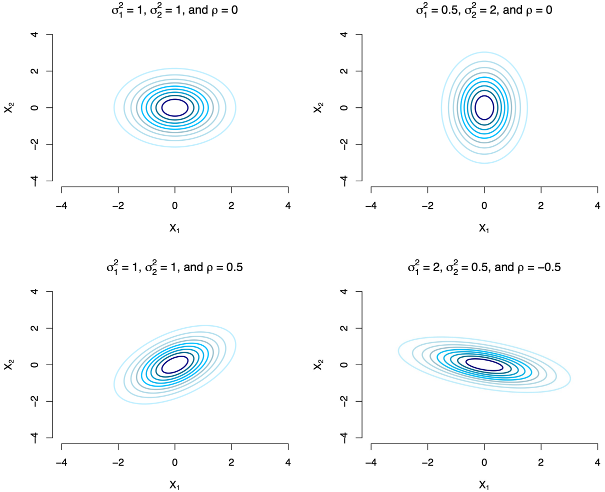

Figure 1.1: Contours of 10^{th}-90^{th} quantiles of the bivariate normal density.

Lemma 1.1: Transforming multivariate normal distributions

Suppose that $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a random length-N vector. Then, for a $K \times N$ constant matrix $\mathbf{C}$, $\mathbf{Cy} \sim \mathcal{N}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal)$.

Lemma 1.1: Proof

- This is provable via Definition 3.
- We know that $\mathbf{y} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$ where $\mathbf{A}\mathbf{A}^\intercal = \boldsymbol{\Sigma}$:

$$\begin{aligned}
\mathbf{C}\mathbf{y} &= \mathbf{C}(\mathbf{A}\mathbf{z} + \boldsymbol{\mu}) \\
&= \mathbf{C}\mathbf{A}\mathbf{z} + \mathbf{C}\boldsymbol{\mu} \\
&\sim \mathcal{N}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\mathbf{A}(\mathbf{C}\mathbf{A})^\intercal) \\
&\sim \mathcal{N}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal)
\end{aligned}$$

- Note that if $\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal$ is singular, then $\mathbf{C}\mathbf{y}$ may have a singular multivariate normal distribution (can be characterized by fewer dimensions).

Lemma 1.2: Univariate-multivariate relationship

$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if $\mathbf{c}^\top \mathbf{y}$ is normally distributed for all $\mathbf{c} \neq \mathbf{0}$.

Lemma 1.2: Proof

- To prove this, we need to show that the first statement implies the second (follows directly from Lemma 1.1), and that the second implies the first. To do this, we'll invoke the Definition 2.
- Suppose that $X = \mathbf{c}^\mathsf{T}\mathbf{y}$ is normally distributed for all nonzero vectors $\mathbf{c}$. Then, for all t, X has MGF given by:

$$M_X(t) = \mathrm{E}[\exp(tX)] = \exp\left((\mathbf{c}^\mathsf{T}\boldsymbol{\mu})t + \frac{1}{2}(\mathbf{c}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{c})t^2 \right).$$

Setting $t = 1$, we have that

$$\mathrm{E}[\exp(X)] = \mathrm{E}[\exp(\mathbf{c}^\mathsf{T}\mathbf{y})] = \exp\left(\mathbf{c}^\mathsf{T}\boldsymbol{\mu} + \frac{1}{2}(\mathbf{c}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{c}) \right),$$

which is nothing other than the multivariate normal MGF (replace the $\mathbf{c}$ with a $\mathbf{t}$ in the above expression and you'll see it).

Lemma 1.3: Orthogonal transformations

Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, and let $\mathbf{Q}$ denote an orthogonal matrix. Then, $\mathbf{Qy} \sim \mathcal{N}(\mathbf{Q}\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Lemma 1.3: Proof

- The proof is straightforward (but you have to remember the properties of orthogonal matrices):

$$\mathbf{Qy} \sim \mathcal{N}(\mathbf{Q}\boldsymbol{\mu}, \mathbf{Q}(\sigma^2\mathbf{I})\mathbf{Q}^\intercal) \equiv \mathcal{N}(\mathbf{Q}\boldsymbol{\mu}, \sigma^2\mathbf{Q}\mathbf{Q}^\intercal) \equiv \mathcal{N}(\mathbf{Q}\boldsymbol{\mu}, \sigma^2\mathbf{I}).$$

- Plain-language meaning: mutually independent random variables with common variance remain mutually independent with common variance under orthogonal transformations.

Example 1.2: Transforming standard normal random variables

- Let X_1, X_2, and X_3 denote i.i.d. standard normal random variables. Further, let

$$
\begin{aligned}
Y_1 &= \frac{1}{\sqrt{3}}(X_1 + X_2 + X_3) \\
Y_2 &= \frac{1}{\sqrt{2}}(X_1 - X_2) \\
Y_3 &= \frac{1}{\sqrt{6}}(X_1 + X_2 - 2X_3).
\end{aligned}
$$

- It is straightforward to show that Y_1, Y_2, and Y_3 are i.i.d. standard normal random variables.
- First, determine the matrix $\mathbf{A}$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$. Then, note that $\mathbf{A}$ is an orthogonal matrix so that $\mathbf{A}^{-1} = \mathbf{A}^\intercal$. Apply Lemma 1.1 and Definition 3 of the multivariate normal distribution (or Lemma 1.3 directly).

Partitioning of the multivariate normal distribution:

- Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We may partition $\mathbf{y}$ as

$$\mathbf{y} \quad = \quad \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix},$$

  where $\mathbf{y}_1$ is of length Q and $\mathbf{y}_2$ of length $N - Q$. The mean vector and covariance matrix can be partitioned accordingly:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}[\mathbf{y}_1] & \mathrm{Cov}[\mathbf{y}_1, \mathbf{y}_2] \\ \mathrm{Cov}[\mathbf{y}_2, \mathbf{y}_1] & \mathrm{Cov}[\mathbf{y}_2] \end{bmatrix}.$$

The multivariate normal distribution

Partitioning: Marginal and conditional distributions

- $\mathbf{y}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{y}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.
  - To show this for $\mathbf{y}_1$, let $\mathbf{y}_1 = \mathbf{C}\mathbf{y}$, where $\mathbf{C} = (\mathbf{I}_{Q\times Q}, \mathbf{0}_{Q\times(N-Q)})$, and recall that $\mathbf{C}\mathbf{y} \sim \mathcal{N}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal)$.
- $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent if and only if $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\intercal = \mathbf{0}$.
  - This can be shown using moment generating functions.
- If $\boldsymbol{\Sigma} \succ 0$, then:

$$\mathbf{y}_1|\mathbf{y}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Theorem 1.2: Distribution of quadratic forms

If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} \succ 0$, then $(\mathbf{y} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2_N$.

Theorem 1.2: Proof

- Since $\boldsymbol{\Sigma} \succ 0$, there is a nonsingular $\mathbf{A}$ with $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$. Indeed, $\mathbf{y} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus,

$$\begin{aligned}
(\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) &= (\mathbf{y} - \boldsymbol{\mu})^\top (\mathbf{A}\mathbf{A}^\top)^{-1}(\mathbf{y} - \boldsymbol{\mu}) \\
&= (\mathbf{y} - \boldsymbol{\mu})^\top (\mathbf{A}^\top)^{-1}\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \\
&= (\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}))^\top (\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})) \\
&= \mathbf{z}^\top \mathbf{z} \sim \chi^2_N.
\end{aligned}$$

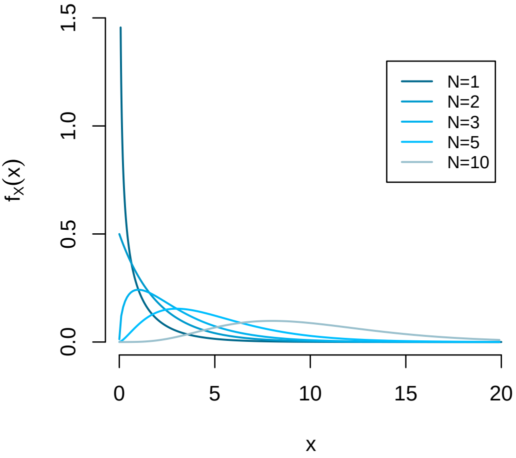

Figure 1.2: The χ^2_N density.

The χ_N^2 distribution: Suppose $X \sim \chi_N^2$

- $X \sim \text{Gamma}(k = N/2, \theta = 2)$.
- $\text{E}[X] = N$.
- $\text{Var}[X] = 2N$.
- X has MGF given by $M_X(t) = (1 - 2t)^{-N/2}$ for $t < 1/2$.

Example 1.3: The eigenvalues are everywhere!

- Suppose $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- Let $Z = 2(Y_1Y_2 - Y_2Y_3 - Y_1Y_3)$.
- How might we derive the MGF, $M_Z(t)$, of Z?

Example 1.3: The eigenvalues are everywhere!

- The first trick is to recognize that Z is indeed a quadratic form.
- $Z = \mathbf{y}^\top \mathbf{A} \mathbf{y}$ for a *symmetric* matrix, $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}.$$

- $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$ by symmetry, and now note that $Z = \mathbf{y}^\top \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top \mathbf{y} = \mathbf{x}^\top \boldsymbol{\Lambda} \mathbf{x}$, where $\mathbf{x} = \mathbf{Q}^\top \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ by Lemma 1.3.
- $\mathbf{A}$ has eigenvalues 2, -1, and -1 (use R), so $Z = 2X_1^2 - X_2^2 - X_3^2$, where X_1^2, X_2^2, and X_3^2 are three i.i.d. χ_1^2 random variables.
- $M_Z(t) = M(2t) \times M(-t) \times M(-t)$, where $M(t)$ is the MGF for a χ_1^2 random variable. $M_Z(t) = (1-4t)^{-1/2}(1+2t)^{-1}$.
- You should be able to come up with all kinds of quadratic forms in $\mathbf{y}$ that have the same MGF. Why?

Theorem 1.3: Distribution of quadratic forms: idempotent case

If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, and let $\mathbf{P}$ be an $N \times N$ symmetric matrix of rank R. Then $(\mathbf{y} - \boldsymbol{\mu})^\intercal\mathbf{P}(\mathbf{y} - \boldsymbol{\mu})/\sigma^2 \sim \chi^2_R$ if and only if $\mathbf{P}$ is idempotent (and hence a projection matrix).

Theorem 1.3: Proof

- Since $\mathbf{P}$ is symmetric, $\mathbf{P}$ is idempotent and rank R if and only if there is an orthogonal $\mathbf{Q}$ such that $\mathbf{Q}^\top \mathbf{P}\mathbf{Q} = \boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda} = \text{diag}(\mathbf{1}_R, \mathbf{0}_{N-R})$.
- Now, let $\mathbf{z} = \mathbf{Q}^\top(\mathbf{y} - \boldsymbol{\mu})$, so that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Therefore,

$$\begin{aligned}
\frac{1}{\sigma^2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{P}(\mathbf{y} - \boldsymbol{\mu}) &= \frac{1}{\sigma^2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top(\mathbf{y} - \boldsymbol{\mu}) \\
&= \frac{1}{\sigma^2}\mathbf{z}^\top \boldsymbol{\Lambda}\mathbf{z} \\
&= \sum_{i=1}^{R}\left(\frac{Z_i}{\sigma}\right)^2 \sim \chi^2_R.
\end{aligned}$$

Theorem 1.4: Independence of $\mathbf{y}^\intercal\mathbf{A}\mathbf{y}$ and $\mathbf{B}\mathbf{y}$

If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{A}$ and $\mathbf{B}$ be constant matrices. Then $\mathbf{y}^\intercal\mathbf{A}\mathbf{y}$ and $\mathbf{B}\mathbf{y}$ are independent if and only if $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$.

Example 1.4: Independence of sample mean and sample variance

- You can use this theorem to show that the sample mean and sample variance of samples from a normal distribution are independent!
- Specifically, suppose $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. Let:

$$\begin{aligned} \mathbf{B} &= (\mathbf{1}_N)^\mathsf{T}/N, \text{ so that } \mathbf{By} = \overline{Y}, \text{ and} \\ \mathbf{A} &= (\mathbf{I} - (\mathbf{J}_N)/N)/(N-1), \text{ so that } \mathbf{y}^\mathsf{T}\mathbf{Ay} = S^2. \end{aligned}$$

- Then, it is straightforward to verify that $\mathbf{B\Sigma A} = \mathbf{0}$.

Another useful fact:

- Using properties of orthogonal transformations of normal random variables can further help you show the following fact:

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{N-1}.$$

- Keep in mind that this is for normally distributed variables.
- We will later prove this finding for more general regression models with normal errors.

Theorem 1.5: Independence of $\mathbf{y}^\intercal\mathbf{P}_1\mathbf{y}$ and $\mathbf{y}^\intercal\mathbf{P}_2\mathbf{y}$

If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{P}_1$ and $\mathbf{P}_2$ be projection matrices. Then, $\mathbf{y}^\intercal\mathbf{P}_1\mathbf{y}$ and $\mathbf{y}^\intercal\mathbf{P}_2\mathbf{y}$ are independent if and only if $\mathbf{P}_2\boldsymbol{\Sigma}\mathbf{P}_1 = \mathbf{0}$.

- This result actually holds for general symmetric matrices, though this version is sufficient for some results we'll later determine. Specifically, this theorem will be useful for us when we talk about hypothesis testing, so we can file this piece of information away for the time being.

Lemma 1.4: Difference of projection matrices

Suppose $\mathbf{P}_1$ and $\mathbf{P}_2$ are projection matrices. If $\mathbf{P}_1 - \mathbf{P}_2 \succcurlyeq 0$, then:

- $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_2\mathbf{P}_1 = \mathbf{P}_2$.
- $\mathbf{P}_1 - \mathbf{P}_2$ is a projection matrix.

- $\mathbf{P}_1$ projects onto a linear space (call it Ω).
- $\mathbf{P}_2$ projects onto a subspace, $\omega \subset \Omega$.
- $\mathbf{P}_1 - \mathbf{P}_2$ projects onto the orthogonal complement of ω within Ω.

Theorem 1.6: $\boldsymbol{\chi}^2$ distribution and projection matrices

Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ be length-N, and $Q_i = (\mathbf{y} - \boldsymbol{\mu})^\intercal \mathbf{P}_i(\mathbf{y} - \boldsymbol{\mu})/\sigma^2$, where $\mathbf{P}_1$ and $\mathbf{P}_2$ are symmetric $N \times N$ matrices. If $Q_i \sim \chi^2_{R_i}$ and $Q_1 - Q_2 \geq 0$, then $Q_1 - Q_2 \sim \chi^2_{R_1 - R_2}$.

Theorem 1.6: Proof

- $Q_i \sim \chi^2_{R_i} \Rightarrow \mathbf{P}_i$ is a projection matrix by Theorem 1.3.
- Because $Q_1 - Q_2 \geq 0$, we know that $\mathbf{P}_1 - \mathbf{P}_2 \succcurlyeq 0$, and so $\mathbf{P}_1 - \mathbf{P}_2$ is a projection matrix by Lemma 1.4.
- Theorem 1.3 further implies that:

$$Q_1 - Q_2 = (\mathbf{y} - \boldsymbol{\mu})^\intercal (\mathbf{P}_1 - \mathbf{P}_2)(\mathbf{y} - \boldsymbol{\mu})/\sigma^2 \sim \chi^2_R,$$

  where $R = \text{rank}(\mathbf{P}_1 - \mathbf{P}_2) = \text{rank}(\mathbf{P}_1) - \text{rank}(\mathbf{P}_2) = R_1 - R_2$.
  - Is this always true, or is there something special about this case? =)
- Note that we find from Theorem 1.5, in addition, that $Q_2 \perp\!\!\!\perp Q_1 - Q_2$, since $\mathbf{P}_2(\mathbf{P}_1 - \mathbf{P}_2) = \mathbf{0}$.

Table of Contents

1 Random vectors/matrices and their properties

2 The multivariate normal distribution

3 Other important distributions

The F-distribution:

- Suppose $U_1 \sim \chi^2_{d_1}$ and $U_2 \sim \chi^2_{d_2}$ are independent. Then,

$$\frac{U_1/d_1}{U_2/d_2} \sim \mathcal{F}_{d_1,d_2},$$

  which we refer to as an F-distribution with d_1 and d_2 degrees of freedom.
- The F-distribution won't make much of an appearance until we talk about hypothesis testing, so we can file this piece of information away for the time being.

Other important distributions

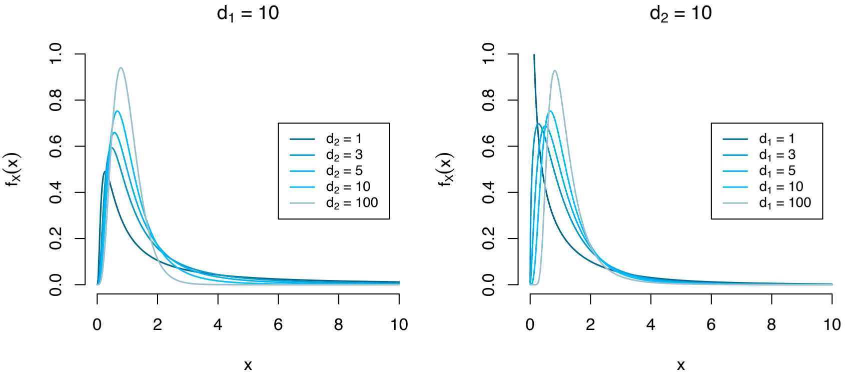

Figure 1.3: The F-distribution.

The non-central χ^2 distribution:

- The non-central χ^2 distribution with N degrees of freedom and non-centrality parameter λ, denoted $\chi^2_N(\lambda)$, has MGF:

$$M_X(t) = (1-2t)^{-N/2} \exp\left(\lambda\left(\frac{1}{1-2t} - 1\right)\right), t < 1/2,$$

  which may differ from some other parameterizations you see.
- The non-central χ^2 distribution can be motivated by the quantity $\sum_{i=1}^N W_i^2$, where $W_1, \ldots, W_N \sim \mathcal{N}(\mu_i, 1)$ are independent; the non-centrality parameter is given by $\lambda = \sum_{i=1}^N \mu_i^2/2$.
  - If $\mu_i = \mu$ for $i = 1, \ldots, N$, then $\lambda = N\mu^2/2$.

Other important distributions

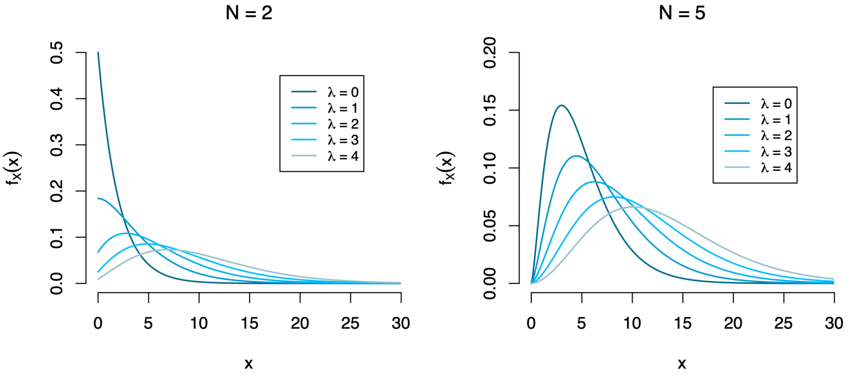

Figure 1.4: The non-central χ^2 distribution.

The non-central χ^2 distribution:

- If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_N, \mathbf{I}_{N\times N})$, then $\mathbf{y}^\intercal\mathbf{y}$ has MGF given by:

$$M_{\mathbf{y}^\intercal\mathbf{y}}(t) = (1-2t)^{-N/2} \exp\left(\frac{\boldsymbol{\mu}^\intercal\boldsymbol{\mu}}{2}\left(\frac{1}{1-2t} - 1\right)\right), t < 1/2.$$

The non-central χ^2 distribution: Properties and results

- If $X \sim \chi^2_N(\lambda)$, then $\text{E}[X] = N + 2\lambda$ and $\text{Var}[X] = 2N + 8\lambda$.
  - These are easily shown using the cumulant-generating function:

$$\begin{aligned}
\text{E}[X] &= \frac{d}{dt} \log M_X(t)\bigg|_{t=0}, \quad \text{and} \\
\text{Var}[X] &= \frac{d^2}{dt^2} \log M_X(t)\bigg|_{t=0}.
\end{aligned}$$

- If $X \sim \chi^2_N(\lambda)$, then $\text{E}[X^{-1}] = 1/(N-2)$ for $N > 2$.

Theorem 1.7: The non-central $\boldsymbol{\chi}^2$ and projection matrices

Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_N, \sigma^2 \mathbf{I}_{N\times N})$ and let $\mathbf{P}$ be a symmetric matrix of rank R. Then, $\mathbf{P}$ is idempotent (and hence an orthogonal projection matrix) if and only if:

$$\frac{1}{\sigma^2}\mathbf{y}^\intercal \mathbf{P}\mathbf{y} \sim \chi^2_R(\boldsymbol{\mu}^\intercal \mathbf{P}\boldsymbol{\mu}/2\sigma^2).$$

The non-central *F*-distribution:

- Suppose $U_1 \sim \chi^2_{d_1}(\lambda)$ and $U_2 \sim \chi^2_{d_2}$ are independent random variables. Then,

$$\frac{U_1/d_1}{U_2/d_2} \sim \mathcal{F}_{d_1,d_2}(\lambda),$$

  which we refer to as a non-central F-distribution with d_1 and d_2 degrees of freedom and non-centrality parameter λ.
- The non-central F-distribution won't make much of an appearance until we talk about hypothesis testing, so we can file this piece of information away for the time being.

Other important distributions

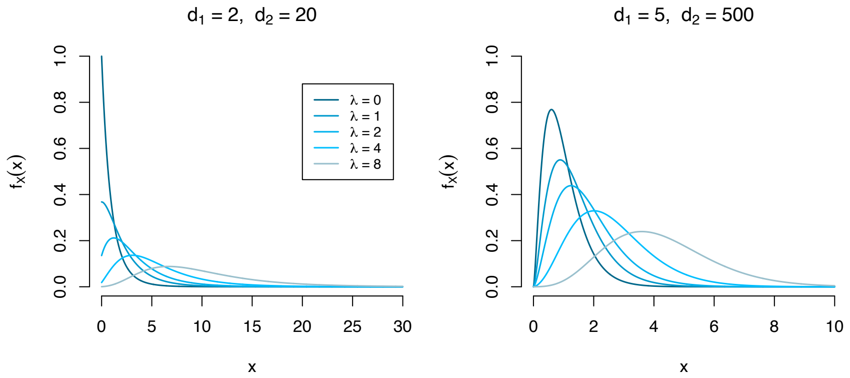

Figure 1.5: The non-central F-distribution.

Summary

This unit:

- Random vectors and random matrices.
- Expectation, variance, and covariance.
- The multivariate normal distribution.
- The χ^2 distribution (central and non-central).
- The F-distribution (central and non-central).

Summary: So far

- Random vectors and matrices; multivariate normal theory.

Summary: Coming up

- Ordinary least squares.
- Hypothesis testing and ANOVA.
- Weighted least squares.
- Confidence regions and prediction.
- Diagnostics.
- Exponential families.
- Generalized linear models.
- Sandwich and bootstrap.
- Quasi-likelihood.
- Hypothesis testing for GLMs.
- Diagnostics for GLMs.
- Further considerations for binary outcomes.
- Nonlinear least squares.