Expectations, Variances, and Distributions of Random Vectors

Supplement to Casella and Berger Section 4.6

A note on notation

Typically, we denote a random variable by an uppercase letter (e.g., X, Y, Z), and a specific realization of that variable by a lowercase letter (e.g., x, y, z). However, when we move to random vectors, this convention conflicts with the notational convention in linear algebra of using lowercase letters for vectors (e.g., $\vec{x}$ or $\mathbf{x}$), and uppercase letters for matrices (e.g., $\mathbf{X}$). Thus, some textbooks adopt the lowercase notation for random vectors from linear algebra. We will adopt this convention here.

Note that Casella and Berger, on the other hand, continue to use uppercase letters for random vectors and lowercase letters for realizations of those random vectors, but in boldface.

All vectors will be assumed to be column vectors, unless otherwise specified.

Expectation of random vectors

Let $y_1, y_2, \ldots, y_n$ be a set of random variables, and define the random vector $\mathbf{y}$ as

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Note that Casella and Berger leave random vectors in vector notation, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, rather than matrix notation, but matrix notation will be more clear for our purposes.

Definition 1. Suppose that $E(y_i) = \mu_i$ for $i = 1, \ldots, n$, and define the $n \times 1$ vector $\boldsymbol{\mu} = \{\mu_i\}$, where the set notation denotes the elements of the vector, i.e.,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}.$$

Then the **expected value of y** is defined as the vector of expected values of its elements:

$$E(\mathbf{y}) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}.$$

Result 1. If $\mathbf{g}(\mathbf{y})$ is a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$, then $E(\mathbf{g}(\mathbf{y})) = \mathbf{g}(E(\mathbf{y}))$.

For example, if $\mathbf{A} : p \times n$, $\mathbf{b} : r \times 1$, and $\mathbf{C} : p \times r$ are matrices of constants, then

$$E(\mathbf{A}\mathbf{y}\mathbf{b}' + \mathbf{C}) = \mathbf{A}E(\mathbf{y})\mathbf{b}' + \mathbf{C} = \mathbf{A}\boldsymbol{\mu}\mathbf{b}' + \mathbf{C}.$$

Variance and covariance of random vectors

Let $\mathbf{x}$ be an $n \times 1$ random vector, and let $\mathbf{y}$ be an $r \times 1$ random vector.

Definition 2. The **covariance of x and y** is

$$Cov(\mathbf{x}, \mathbf{y}) = E\left[(\mathbf{x} - E(\mathbf{x}))(\mathbf{y} - E(\mathbf{y}))'\right].$$

Note that the matrix $Cov(\mathbf{x}, \mathbf{y})$ has dimension $n \times r$, and the $(i, j)^{th}$ element is $Cov(x_i, y_j)$.

Definition 3. The **variance of x** is the $n \times n$ matrix

$$Var(\mathbf{x}) = Cov(\mathbf{x}, \mathbf{x}) = E\left[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))'\right].$$

Result 2. $Var(\mathbf{x})$ is a symmetric non-negative definite matrix.

As with the variance and covariance of random variables, there are "shortcut" formulas to finding the variance and covariance of random vectors.

Result 3. $Cov(\mathbf{x}, \mathbf{y}) = E(\mathbf{x}\mathbf{y}') - E(\mathbf{x})E(\mathbf{y})'$.

Result 4. $Var(\mathbf{x}) = E(\mathbf{x}\mathbf{x}') - E(\mathbf{x})E(\mathbf{x})'$.

The following properties follow from the definitions of expectation and covariance of random vectors, and from properties of matrices.

Result 5. For any constant matrices $\mathbf{A} : m \times n$ and $\mathbf{B} : p \times r$,

a. $Cov(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}) = \mathbf{A}Cov(\mathbf{x}, \mathbf{y})\mathbf{B}'$

b. $Var(\mathbf{A}\mathbf{x}) = Cov(\mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x}) = \mathbf{A}Var(\mathbf{x})\mathbf{A}'$

Multivariate normal (MVN) distribution

Definition 4. Suppose that random vector $\mathbf{y} : n \times 1$ with support $\mathbb{R}^n$ has joint probability density function,

$$f(\mathbf{y}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}},$$

for $\boldsymbol{\mu} : n \times 1 \in \mathbb{R}^n$ and positive definite matrix $\boldsymbol{\Sigma} : n \times n$. Then $\mathbf{y}$ is said to have a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, denoted by $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ and $\mathbf{I}_n$ is an $n \times n$ identity matrix, then the pdf simplifies to

$$f(\mathbf{y}) = \frac{\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu})\right\}}{(2\pi\sigma^2)^{\frac{n}{2}}} = \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\right\}}{(2\pi\sigma^2)^{\frac{n}{2}}}.$$

Properties of the MVN distribution

The following properties hold for the multivariate normal distribution.

Result 6. Suppose $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then the following results can be established.

a. Moment generating function:

$$M_{\mathbf{y}}(t) = E[\exp(\mathbf{t}'\mathbf{y})] = \exp\left\{\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right\}.$$

b. $E(\mathbf{y}) = \boldsymbol{\mu}$.

c. $Var(\mathbf{y}) = \mathbf{\Sigma}$.

d. If $\mathbf{A}$ is an $r \times n$ matrix of constants, then

$$\mathbf{Ay} \sim N_r(\mathbf{A}\boldsymbol{\mu}, \mathbf{A\Sigma A}').$$

e. Let $\mathbf{y}$ be an $n \times 1$ random vector with distribution $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$. The MVN density function is constant for all $\mathbf{y}$ that satisfy

$$(\mathbf{y} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = c.$$

The above equation is the equation of an n-dimensional ellipsoid.

The MVN distribution and independence

In general, if two random vectors $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated (i.e., $Cov(\mathbf{x}, \mathbf{y}) = 0$), we *cannot* assume that they are independent (though the reverse always holds). However, if the two vectors form a multivariate normal distribution, zero covariance will imply independence, as stated by the following result.

Result 7. Suppose that $\mathbf{y}$ is a $(p+q) \times 1$ random vector with distribution $\mathbf{y} \sim N_{p+q}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{pmatrix} = \mathbf{\Sigma}_1 \oplus \mathbf{\Sigma}_2,$$

where $\mathbf{\Sigma}_1$ is $p \times p$ and $\mathbf{\Sigma}_2$ is $q \times q$. If we partition $\mathbf{y}$ as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$$

with $\mathbf{y}_1 : p \times 1$ and $\mathbf{y}_2 : q \times 1$, then $\mathbf{y}_1$ is independent of $\mathbf{y}_2$. This result says that if two random vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ are uncorrelated and have a joint multivariate normal distribution, then the random vectors are independent.

The *joint* multivariate distribution is crucial to this result, however. Maybe people mistakenly think that if $\mathbf{x} \sim N_n(\boldsymbol{\mu}_x, \mathbf{\Sigma}_x)$ and $\mathbf{y} \sim N_n(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y)$ with $Cov(\mathbf{x}, \mathbf{y}) = 0$, then the two random vectors are independent, but *that is not necessarily the case*! (See, for example, Rosenthal's "Rant About Uncorrelated Normal Random Variables"[1].)

On the other hand, if two normal random vectors are *independent*, then their joint distribution is multivariate normal.

Result 8. Let $\mathbf{y}_1 \sim N_p(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$ and $\mathbf{y}_2 \sim N_q(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2)$ where $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent. Then

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{pmatrix} \right).$$

Conditional MVN distributions

Let $\mathbf{y}$ be a $(p+q) \times 1$ random vector and denote a realization of the random vector by $\ddot{\mathbf{y}}$. Suppose that $\mathbf{y}$ is distributed as $\mathbf{y} \sim N_{p+q}(\boldsymbol{\mu}, \mathbf{\Sigma})$. Partition $\mathbf{y}$ as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$$

with $\mathbf{y}_1 : p \times 1$ and $\mathbf{y}_2 : q \times 1$. Partition $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ conformably, as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

[1] https://probability.ca/jeff/teaching/uncornor.html

We often may be interested in the conditional distribution of $\mathbf{y}_1$ given a particular value of $\mathbf{y}_2$. This is expressed in the following result.

Result 9. If $\boldsymbol{\Sigma}_{22}$ is positive definite, then conditional on $\mathbf{y}_2 = \ddot{\mathbf{y}}$, $\mathbf{y}_1$ has the multivariate normal distribution

$$\mathbf{y}_1 | \mathbf{y}_2 = \ddot{\mathbf{y}} \sim N_p(\boldsymbol{\mu}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2}),$$

where

$$\boldsymbol{\mu}_{1\cdot 2} := \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\ddot{\mathbf{y}}_2 - \boldsymbol{\mu}_2),$$

and

$$\boldsymbol{\Sigma}_{11\cdot 2} := \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

The previous result took a multivariate random vector, partitioned it into two vectors, and gave the conditional distribution of one of these vectors given the other. But what if we are given the conditional distribution and want to go back to the joint distribution?

Result 10. Suppose $\mathbf{v}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{v}_2 | \mathbf{v}_1 \sim N_{p-r}(\mathbf{A}\mathbf{v}_1 + \mathbf{b}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ does not depend on $\mathbf{v}_1$. Then

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\mu = \begin{pmatrix} \boldsymbol{\mu}_1 \\ A\boldsymbol{\mu}_1 + \mathbf{b} \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{11}\mathbf{A}' \\ \mathbf{A}\boldsymbol{\Sigma} & \boldsymbol{\Omega} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' \end{pmatrix}.$$

MVN distribution's relation to the chi-squared distribution

Given an $n \times 1$ random vector $\mathbf{x}$ and $n \times n$ constant matrix $\mathbf{A}$, the random variable $Q = \mathbf{x}'\mathbf{A}\mathbf{x}$ is called a **quadratic form in x**. We will further examine distributions of quadratic forms in Stat 502[2], but for now, we discuss one specific example.

Definition 5. Suppose $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \mathbf{I}_n)$. Define the random variable $Q = \mathbf{x}'\mathbf{x}$. Then Q is said to have a **noncentral chi-squared distribution** with n degrees of freedom and noncentrality parameter $\lambda = \boldsymbol{\mu}'\boldsymbol{\mu}/2$. This is denoted by $Q \sim \chi^2(n, \lambda)$.

If $\boldsymbol{\mu} = \mathbf{0}$, then $\lambda = 0$, and Q is said to have a **central chi-squared distribution** with n degrees of freedom, denoted by $Q \sim \chi^2(n)$.

The following results summarize some basic properties of noncentral chi-squared distributions.

Result 11. Let $Q \sim \chi^2_{n,\lambda}$. Then the following properties hold.

a. Probability density function:

$$f_Q(q) = \sum_{j=0}^{\infty} \frac{e^{\lambda}\lambda^j}{j!} f_{n+2j}(q),$$

where $f_{n+2j}(q)$ is the density function for a central chi-squared random variable having $n + 2j$ degrees of freedom. That is,

$$f_i(q) = \frac{e^{q/2}q^{i/2-1}}{2^{i/2}\Gamma(i/2)}.$$

b. Moment generating function:

$$M_Q(t) = (1-2t)^{-n/2}\exp[2t\lambda/(1-2t)].$$

c. $E(Q) = n + 2\lambda$.

d. $Var(Q) = 2n + 8\lambda$.

[2]It turns out that the sample variance can be expressed as a quadratic form.

e. If $Q_i \overset{ind}{\sim} \chi^2(\nu_i, \lambda_i)$ for $i = 1, \ldots, k$, then

$$\sum_{i=1}^{k} Q_i \sim \chi^2\left(\sum_{i=1}^{k} \nu_i, \sum_{i=1}^{k} \lambda_i\right).$$

Application: Linear models

Consider a response vector $\mathbf{y} = (y_1, \ldots, y_n)'$, where each y_i is a variable of interest measured on observational unit i in a sample of size n. We would like to predict the response for future observations using a set of predictor variables. If $\mathbf{y}$ follows a **linear model**, we can write it in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is an $n \times p$ matrix of known constants (information from the predictor variables), $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters that we would like to estimate, and $\boldsymbol{\epsilon}$ is a random error vector with expectation $\mathbf{0}$ (an $n \times 1$ vector of zeroes) and variance-covariance matrix $\boldsymbol{\Sigma}$ (an $n \times n$ matrix).

Example: Multiple linear regression An investigator obtained a random sample of n incoming MSU undergraduates. On each case, the investigator observed the undergraduate's high school (HS) GPA and the number of math courses taken in HS. After their first year, the investigator also obtained their college GPA. The investigator believes that college GPA can be predicted from HS GPA as well as the number of HS math courses. Denote the college GPA for the ith case as y_i, the HS GPA for the ith case as g_i, and the number of HS math courses as m_i. Then a multiple linear regression model for this scenario is

$$y_i = \beta_0 + \beta_1 g_i + \beta_2 m_i + \epsilon_i,$$

where β_0 is the intercept term, β_1 is the regression coefficient for HS GPA, and β_2 is the regression coefficient for number of HS math courses. We also assume that

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

which implies that

$$y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 g_i + \beta_2 m_i, \sigma^2).$$

Express this model in matrix terms.

Example: One mean Consider the linear model $y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, and define $\mathbf{y} = (y_1, \ldots, y_n)'$. Two statistics of primary interest are the sample mean and sample variance, respectively:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \text{ and } S^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}.$$

Express $\bar{y}$ and S^2 as linear and quadratic functions of $\mathbf{y}$.

The previous example can be extended to the **general linear model**,

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Omega}),$$

where $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, σ^2 is a scalar, and $\boldsymbol{\Omega}$ is a $n \times n$ positive definite matrix. The **generalized least squares (and maximum likelihood) estimator** of $\boldsymbol{\beta}$ is the minimizer of the **sum of squared errors**,

$$SSE(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

It turns out that this minimizer is a linear function of $\mathbf{y}$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y},$$

which also implies that the predicted values are a linear function of $\mathbf{y}$,

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}.$$

The matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}$ is called a **projection operator**—it projects the response vector onto the column space of $\mathbf{X}$ (all vectors in $\mathbb{R}^n$ that can be expressed in the form $\mathbf{Xb}$ for some $\mathbf{b} \in \mathbb{R}^p$).

Additionally, an unbiased estimator of σ^2 is $SSE(\hat{\boldsymbol{\beta}})/(n-p)$, which can be expressed as a quadratic function of $\mathbf{y}$:

$$\hat{\sigma^2} = \frac{SSE(\hat{\boldsymbol{\beta}})}{n-p} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{\mathbf{e}'\boldsymbol{\Omega}^{-1}\mathbf{e}}{n-p} = \frac{\mathbf{y}'\boldsymbol{\Omega}^{-1}(\mathbf{I}_n - \mathbf{P})\mathbf{y}}{n-p},$$

where $\mathbf{e}$ is the $n \times 1$ vector of residuals.

References

Much of this material was taken directly from Dr. Robert Boik's Stat 505 Lecture Notes, *A Pair of Primers: Primer on Matrix Analysis and Primer on Linear Statistical Models*, Fall 2007 edition. Also referenced was the book *Matrix Algebra from a Statistician's Perspective* by David A. Harville, Springer, 1997.