Refresher on Probability and Matrix Operations

Charlie Gibbons
Political Science 236

September 9, 2008

Outline

Matrix operations

Here are some of the important properties of matrix and vector operations. You don't need to memorize them, but you should be able to apply them to questions on your problem sets.

Matrix operations

$$
\begin{array}{rcl}
A + B &=& B + A \\
A(B + C) &=& AB + AC \\
(A + B)C &=& AC + BC \\
AB &\neq& BA \qquad \text{(in general)} \\
AB &=& AC \quad \not\Rightarrow \quad B = C \\
AB &=& 0 \quad \not\Rightarrow \quad A = 0 \text{ or } B = 0 \\
(A + B)' &=& A' + B' \\
(ABC)' &=& C'(AB)' = C'B'A' \\
\left(M^{-1}\right)' &=& \left(M'\right)^{-1} \\
(LMN)^{-1} &=& N^{-1}(LM)^{-1} = N^{-1}M^{-1}L^{-1}
\end{array}
$$

Matrix operations

What's different between the matrices A, B, and C versus L, M, and N?

Matrix operations

What's different between the matrices A, B, and C versus L, M, and N?

Since we are taking the inverses of the latter group, they must be *square* and *full rank*. This is known as being *non-singular*.

Matrix operations

What's different between the matrices A, B, and C versus L, M, and N?

Since we are taking the inverses of the latter group, they must be *square* and *full rank*. This is known as being *non-singular*.

Think of the matrix as a system of equations. each row in M is a set of coefficients for the values of a vector of unknown variables, say r, and solutions, s. If the matrix is invertible, you can solve $Ar = s$.

Matrix operations

What's different between the matrices A, B, and C versus L, M, and N?

Since we are taking the inverses of the latter group, they must be *square* and *full rank*. This is known as being *non-singular*.

Think of the matrix as a system of equations. each row in M is a set of coefficients for the values of a vector of unknown variables, say r, and solutions, s. If the matrix is invertible, you can solve $Ar = s$.

In the context of regression, the M matrix is our observed covariates (usually called X), s is a vector of outcomes (y), and r is the vector of coefficients that we are trying to find (β).

Vectors and summation

Define l as a conformable vector of ones.

Vectors and summation

Define l as a conformable vector of ones.

We can express a sum in terms of vector multiplication:

$$\sum_{i=1}^{n} x_i = x'l$$

Vectors and summation

Define l as a conformable vector of ones.

We can express a sum in terms of vector multiplication:

$$\sum_{i=1}^{n} x_i = x'l$$

We can do the same for a sum of squares of x:

$$\sum_{i=1}^{n} x_i^2 = x'x$$

Preliminary definitions

The *sample space* Ω is the set of all possible outcomes of our "experiment."

Note that *experiment* has the meaning in probability theory of being any situation in which the final outcome is unknown and is distinct from the way that we will define an experiment in class.

Preliminary definitions

The *sample space* Ω is the set of all possible outcomes of our "experiment."

Note that *experiment* has the meaning in probability theory of being any situation in which the final outcome is unknown and is distinct from the way that we will define an experiment in class.

An *event* ω is a collection of possible realizations from the sample space.

Preliminary definitions

The *sample space* Ω is the set of all possible outcomes of our "experiment."

Note that *experiment* has the meaning in probability theory of being any situation in which the final outcome is unknown and is distinct from the way that we will define an experiment in class.

An *event* ω is a collection of possible realizations from the sample space.

Events are *disjoint* if they do not share a common element; *i.e.*, $\omega_i \cap \omega_j = \varnothing$.

Preliminary definitions

The *sample space* Ω is the set of all possible outcomes of our "experiment."

Note that *experiment* has the meaning in probability theory of being any situation in which the final outcome is unknown and is distinct from the way that we will define an experiment in class.

An *event* ω is a collection of possible realizations from the sample space.

Events are *disjoint* if they do not share a common element; *i.e.*, $\omega_i \cap \omega_j = \varnothing$.

An *elementary event* is an event that only contains a single realization from the sample space.

Preliminary definitions

Lastly, though getting a bit ahead, two events are statistically independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

Calculating probabilities

The classical definition of probability states that, for a sample space containing equally-likely elementary events, then the probability of an event ω is the ratio of the number of elementary events in ω to that of Ω; *i.e.*,

$$P(\omega) = \frac{\#(\omega)}{\#(\Omega)}$$

Calculating probabilities

The axiomatic definition of probability defines probability by stating that

$$
\begin{array}{rcl}
\Pr(\omega_i) & \geq & 0 \quad \forall\, \omega_i \in \Omega \\
\Pr(\Omega) & = & 1 \\
\Pr(\omega_1 \cup \omega_2 \cup \cdots \cup \omega_n) & = & \Pr(\omega_1) + \Pr(\omega_2) + \cdots + \Pr(\omega_n) \\
 & & \text{for pairwise disjoint } \omega_1, \ldots, \omega_n
\end{array}
$$

Probability properties

Let A and B be events in the sample space Ω.

$$
\begin{aligned}
\Pr(\Omega) &= 1 \\
\Pr(\varnothing) &= 0 \\
\Pr(A) &\geq 0 \\
\Pr(A^c) &= 1 - \Pr(A) \\
\Pr(A) &\leq \Pr(B) \quad \forall\, A \subseteq B \\
\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\
\Pr(A \cup B) &\leq \Pr(A) + \Pr(B) \qquad \text{(Boole's inequality)} \\
\Pr(A \cap B) &\geq \Pr(A) + \Pr(B) - 1 \qquad \text{(Bonferroni's inequality)}
\end{aligned}
$$

Probability formulae

The *conditional probability* of an event A given that an event B has occurred is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Probability formulae

The *conditional probability* of an event A given that an event B has occurred is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

This equation reflects that you received additional information about the probability of A knowing that B occurred.

Probability formulae

The *conditional probability* of an event A given that an event B has occurred is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

This equation reflects that you received additional information about the probability of A knowing that B occurred.

$\Pr(B)$ is in the denominator because the sample space has been reduced from the full space Ω to just that portion in which B arises.

Probability formulae

The *law of total probability* holds that, for a countable *partition* of Ω, $\{B_i\}_{i=1}^N$ (*i.e.*, $\cup_{i=1}^N B_i = \Omega$ and $B_j \cap B_k = \varnothing\ \forall j \neq k$), then

$$\Pr(A) = \sum_{i=1}^{N} \Pr(A|B_i) \Pr(B_i)$$

Probability formulae

Bayes' rule states that

$$\begin{aligned}\Pr(B_i|A) &= \frac{\Pr(B_i \cap A)}{\sum_{j=1}^{N} \Pr(B_j \cap A)} = \frac{\Pr(A|B_i)\Pr(B_i)}{\sum_{j=1}^{N} \Pr(A|B_j)\Pr(B_j)} \\ &= \frac{\Pr(A|B_i)\Pr(B_i)}{\Pr(A)}\end{aligned}$$

Probability formulae

Bayes' rule states that

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

for the two event case.

Random variables

A *random variable* is actually a function that maps every outcome in the sample space to the real line. Formally, $X : \Omega \to \mathbb{R}$.

Random variables

A *random variable* is actually a function that maps every outcome in the sample space to the real line. Formally, $X : \Omega \to \mathbb{R}$.

If we want to find the probability of some subset ω of Ω, we can induce a probability onto a random variable X. Let $X(\omega) = x$. Then,

$$\Pr(\omega \in \Omega \,:\, X(\omega) = x) = \Pr(X = x)$$

Random variables

Define the *cumulative distribution function (CDF)*, $F_X(x)$, as $\Pr(X \leq x)$. The CDF has three important properties:

$$\begin{aligned}
\lim_{x \to -\infty} F_X(x) &= 0 \\
\lim_{x \to \infty} F_X(x) &= 1 \\
\frac{dF_X(x)}{dX} &\geq 0 \qquad (i.e., \text{ the CDF is non-decreasing})
\end{aligned}$$

Random variables

A *continuous random variable* has a sample space with an uncountable number of outcomes. Here, the CDF is defined as

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$$

For a discrete random variable, which has a countable number of outcomes, the CDF is defined as

$$F_X(x) = \sum_{y=-\infty}^{x} \Pr(X = y)$$

Random variables

We can define the *probability density function (PDF)* for a continuous variable as

$$f_X(x) = \frac{d}{dX} F_X(x) = f_X(x)$$

by the Fundamental Theorem of Calculus.

It can be defined for a discrete random variable as

$$f_X(x) = \Pr(X = x)$$

Note that $f_X(x) \geq 0 \quad \forall\, x$.

Random variables

We can define the *probability density function (PDF)* for a continuous variable as

$$f_X(x) = \frac{d}{dX} F_X(x) = f_X(x)$$

by the Fundamental Theorem of Calculus.

Joint Probabilities

Previously, we considered the distribution of a lone random variable. Now we will consider the joint distribution of several random variables. For simplicity, we will restrict ourselves to the case of two random variables, but the provided results can easily be extended to higher dimensions.

Joint Probabilities

The *joint cumulative distribution function (joint CDF)*, $F_{X,Y}(x, y)$, of the random variables X and Y is defined by

$$F_{X,Y}(x, y) = \Pr(X \leq x,\ Y \leq y)$$

As with any CDF, $F_{X,Y}(x, y)$ must equal 1 as x and y go to infinity.

Joint Probabilities

The *joint probability mass function (joint PMF)*, $f_{X,Y}$ is defined by

$$F_{X,Y}(x,y) = \sum_{s=-\infty}^{x} \sum_{t=-\infty}^{y} f_{X,Y}(s,t)$$

Joint Probabilities

The *joint probability mass function (joint PMF)*, $f_{X,Y}$ is defined by

$$F_{X,Y}(x,y) = \sum_{s=-\infty}^{x} \sum_{t=-\infty}^{y} f_{X,Y}(s,t)$$

The *joint probability density function (joint PDF)*, $f_{X,Y}$ is defined by

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)\, ds\, dt$$

Joint Probabilities

The *marginal cumulative distribution function (marginal CDF)* of X, $F_X(x)$ is

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y).$$

Joint Probabilities

The *marginal cumulative distribution function (marginal CDF)* of X, $F_X(x)$ is

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y).$$

The *marginal PMF* of X is

$$f_X(x) = \sum_{y=-\infty}^{\infty} f_{X,Y}(x, y)$$

Joint Probabilities

The *marginal cumulative distribution function (marginal CDF)* of X, $F_X(x)$ is

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y).$$

The *marginal PMF* of X is

$$f_X(x) = \sum_{y=-\infty}^{\infty} f_{X,Y}(x, y)$$

The *marginal PDF* of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

You are "integrating out" y from the joint PDF.

Joint Probabilities

Note that, while a marginal PDF (PMF) can be found from a joint PDF (PMF), the converse is not true; there are an infinite number of joint PDFs (PMFs) that could be described by a given marginal PDF (PMF).

Independence

If X and Y are independent, then

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

and

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Conditional Probabilities

The *conditional PDF (PMF) of* Y *given* $X = x$, $f_{Y|X}(y|X = x)$, is defined by

$$f_{X,Y}(x, y) = f_{Y|X}(y|X = x)f_X(x)$$

As for any PDF (PMF), over the support of Y, the conditional PDF (PMF) must integrate (sum) to 1. It must also be non-negative for all real values.

Conditional Probabilities

The *conditional PDF (PMF) of* Y *given* $X = x$, $f_{Y|X}(y|X = x)$, is defined by

$$f_{X,Y}(x, y) = f_{Y|X}(y|X = x)f_X(x)$$

As for any PDF (PMF), over the support of Y, the conditional PDF (PMF) must integrate (sum) to 1. It must also be non-negative for all real values.

For discrete random variables, we see that the conditional PMF is

$$\begin{aligned} f_{Y|X}(y|X = x) &= \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{\Pr(Y = y, X = x)}{\Pr(X = x)} \\ &= \Pr(Y = y|X = x) \end{aligned}$$

Conditional Probabilities

Question: What is random in the conditional distribution of Y, $f_{Y|X}(y|X = x)$?

Conditional Probabilities

If X and Y are independent, then

$$f_{Y|X}(y|X = x) = f_Y(y)$$

Conditional Probabilities

If X and Y are independent, then

$$f_{Y|X}(y|X = x) = f_Y(y)$$

This implies that knowing X gives you no additional ability to predict Y, an intuitive notion underlying independence.