CHAPTER 0

Tensors

0.1 What is a tensor?

0.1.1 *Difficult questions*

In one form or another, the title of this section is a question that I have frequently been asked, often by professional colleagues who should know better, but sometimes by unsuspecting strangers at social gatherings. This is not an ideal conversation starter, but politeness demands an answer. The question is usually asked out of innocent curiosity or out of simple naivety, but I invariably assume that my interlocutor has a burning desire to know the truth, at least as I see it. The question could hardly be more straightforward, and yet I have been woefully unsuccessful at providing anything resembling a satisfactory answer. Even a partial answer that elicited a follow-up question would rate as a success, but this is rare, especially so in mixed company at cocktail parties. Why is it that the simplest questions are so often the hardest to answer? This note explains my current thoughts on the topic, and will serve as a guide should someone be foolish enough to raise the question at a future occasion.

As every successful politician knows, the essential first step in answering any difficult question is to 're-phrase the question.' This can mean anything from 'the question I think was intended' to 'the question I know how to answer' to 'the question I want to be asked.' Even a straightforward, apparently unambiguous, question such as 'what is a tensor?' is susceptible to this ploy. Here are some re-phrased versions of the question, roughly in increasing order of difficulty.

1. What is the mathematical definition of a tensor?
2. Is a tensor a kind of vector?
3. Is a vector a kind of tensor?
4. Is a matrix a special kind of tensor?
5. When you write λ^i, do you mean a row vector or a column vector?
6. What are tensors used for?
7. I know that a vector has magnitude and direction, but what does a tensor look like?
8. How would I recognize a tensor if I met one in the dark?

The first of these questions is easily disposed of by the stock answer, courtesy of vector spaces, 'a tensor is an element of a tensor space.' The second and third questions are answered in a single word, 'yes.' Question 4 is only slightly more difficult, the answer being 'yes and no.' Question 5 deals with typographical strategy: it all depends on whether the elements are written across the page, down the page, or diagonally.

Audiences are invariably nonplussed by the directness and simplicity of these answers. The obvious correctness of the answer to question 1 is eloquent testimony to the persuasive power of modern mathematics. What more is there to say?

Regarding question 7, one could preface a reply by stating that vectors need have neither magnitude nor direction, and the same is true for tensors. But that, I suspect, is more likely to confuse than to enlighten.

Question 8 is the really tough nut. How, for example, can one tell from observation that an electric field is a vector and a magnetic field is a tensor? After all, they do not look very different. In my experience, for these and other reasons, the latter questions are best evaded. Distraction tactics often work well. Offer to extemporize on Aristotelian philosophy. Even volunteer to explain

confidence intervals. On no account be drawn into geometrical descriptions or physical analogies in response to questions 7–8. That way lies certain failure. These, of course, are precisely the questions that every inquisitive student wants to ask, but it is a futile and anti-social exercise to give anything other than a flippant answer, particularly at social gatherings. The more noble the attempt at answering honestly, the more demoralizing the failure, so much so that even an expert begins to have doubts.

0.1.2 *Definitions and/or examples*

With a little training it is fairly easy even for a child to identify a chimpanzee, a tiger or a human being by sight. But to define what is meant by the term 'chimpanzee' in language that would satisfy a logician or mathematician is a much more difficult task. One would not normally cite this inability to produce a definition as evidence that the child does not know what a chimpanzee is. For taxonomic purposes, examples and descriptive characteristics are more useful than definitions.

The purpose of definitions in mathematics is to crystallize essential properties, while eliminating irrelevant subject-matter details. When divorced from examples this is a mindless antiseptic exercise, but when viewed in its proper context a good definition can be unifying and enlightening. A vector, for example, can be described as a quantity with magnitude and direction. This is not satisfactory as a mathematical definition because no mention is made of addition or the triangle rule, but it is nonetheless a useful description in many physical contexts. The standard mathematical definition as an element of a certain kind of set, is cleaner, in many ways more powerful, but is totally lacking in descriptive power and geometrical insight.

In the historical development of science and mathematics, concepts are invariably encountered before adequate definitions are in place. On the whole, the opposite is true in textbooks. However, the lack of a satisfactory definition need not be a barrier to progress. Indeed substantial progress and good understanding are essential in order to formulate good definitions. It can be a useful didactic tool to explore a number of examples before the unifying theme is revealed. The notes in this chapter are neither complete nor self-contained. The treatment of examples is of necessity somewhat condensed. Concepts arise in examples before they are defined. This reversal of the familiar order will confuse some readers, but it may help others to appreciate better the relation between examples and definitions. Even if a satisfactory definition is not achieved, the examples will have served their purpose, and we will have a good working idea of what a tensor is.

0.2 Straightforward answers?

0.2.1 *What is a vector?*

It is assumed that the reader has some familiarity with vectors, yet it is felt necessary to address this question head-on. A vector space is a set $\mathcal{V}$ that is closed under a commutative binary operation called addition, and closed under multiplication by scalars. In symbols,

$$\alpha_1 v_1 + \alpha_2 v_2 = \alpha_2 v_2 + \alpha_1 v_1 \in \mathcal{V}$$

for all $v_1, v_2 \in \mathcal{V}$ and for all scalars α_1, α_2. Unless otherwise specified, scalars are taken to be real numbers, although in many cases complex numbers would serve equally well. A vector is then defined to be an element of a vector space.

The complete definition of a vector space requires a number of additional conditions that are trivially satisfied by all examples in this book, but which need to be checked in more exotic cases. In particular, it is required that there exist a point $0 \in \mathcal{V}$ such that $v + 0 = v$ and $v + (-v) = 0$ for every v. In addition, associative and distributive conditions are needed to ensure that $1v = v$ and $v + v = 2v$ for every v. Full details can be found in Halmos, (1958, section 2).

The vector property is thus a property of a set, not a property of individual elements taken in isolation. According to the definition, therefore, the ordered pair $v = (0.25, 0.75)$ is not a vector

unless and until we specify a suitable set $\mathcal{V}$ to which v belongs. Consider in particular the following sets:

$$\mathcal{V}_1 = \{(\xi_1, \xi_2) : \xi_1, \xi_2 \in \mathcal{R}\};$$
$$\mathcal{V}_2 = \{(\xi_1, \xi_2) : 0 \leq \xi_1, \xi_2 \leq 1\};$$
$$\mathcal{V}_3 = \{(\xi, 3\xi) : \xi \in \mathcal{R}\};$$

where $\mathcal{R}$ denotes the real line. Evidently v is a point in each of these sets, but only $\mathcal{V}_1$ and $\mathcal{V}_3$ are vector spaces with the usual definitions of addition and multiplication by scalars.

The preceding definitions of vector and vector space are entirely abstract. No attempt is made to convey a picture of what it is that constitutes a vector or of what a vector 'looks like' in a physical or operational sense. That failure is understandable, inevitable, and ultimately desirable in view of the enormous variety of objects, even within statistics, that can be considered as vectors. In order to strike a balance, therefore, we make liberal use of examples to supplement the definitions throughout these notes.

0.2.2 *What is a tensor?*

In section 1.5 we define the tensor product of vector spaces, $\mathcal{V}_1$ and $\mathcal{V}_2$. It suffices for the moment to say that the tensor product $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ is a new vector space constructed from the component spaces $\mathcal{V}_1$ and $\mathcal{V}_2$. A tensor of order one is a vector, i.e. an element of a vector space. A tensor of order two is an element of a tensor product space $\mathcal{V}_1 \otimes \mathcal{V}_2$. Tensors of higher order are defined by repeated tensor products of component spaces. Thus all vectors are tensors, and, since the tensor product space is a vector space, all tensors are vectors. This may sound a little strange, even paradoxical, perhaps. In practice, when we say that some quantity is a tensor both the order and the component spaces are usually implied by notation and context. In contexts where tensors of various orders occur, it is common to use the word 'vector' to mean 'tensor of order one,' implying that the space in question is not a tensor product of component spaces.

Even at this stage, although we have not yet defined the tensor product, it is possible to give a simple example to convey the flavour of what is meant by tensor product.

Let $\mathcal{V}$ be the space of real-valued quadratic functions on the interval $[0, 1]$. An element in $\mathcal{V}$ is thus a function v whose numerical value at the point $x \in [0, 1]$ is given by the formula

$$v(x) = \xi_0 + \xi_1 x + \xi_2 x^2,$$

for some real-valued coefficients ξ_0, ξ_1, ξ_2. Although the elements in $\mathcal{V}$ are real-valued functions, they are in 1–1 correspondence with the set of vectors $(\xi_0, \xi_1, \xi_2) \in \mathcal{R}^3$. Evidently $\mathcal{V}$ is a three-dimensional vector space with respect to the usual definitions of addition of polynomials and multiplication of polynomials by scalars. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be independent copies of $\mathcal{V}$. In other words, elements of $\mathcal{V}_1$ are quadratic functions of a variable x_1: elements of $\mathcal{V}_2$ are quadratic functions of a variable x_2. The tensor product space $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ is the space of functions on the unit square, that are quadratic in x_1 for each fixed value of x_2, and quadratic in x_2 for each fixed value of x_1. In symbols,

$$\mathcal{V}_1 = \text{span}\{1,\, x_1,\, x_1^2\}; \qquad \mathcal{V}_2 = \text{span}\{1,\, x_2,\, x_2^2\};$$
$$\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2 = \text{span}\{1,\, x_1,\, x_1^2,\, x_2,\, x_1x_2,\, x_1^2x_2,\, x_2^2,\, x_1x_2^2,\, x_1^2x_2^2\}.$$

Elements in $\mathcal{W}$ are thus linear combinations of formal products of elements in the component spaces. In this example where $\mathcal{V}_1$ and $\mathcal{V}_2$ are essentially (apart from notation) the same space, the tensor product space is often denoted by the symbol $\mathcal{V}^{\otimes 2}$.

Two examples of functions in $\mathcal{W}$ are the following:

$$w_1(x) = 4 - x_1^2 - x_2^2 + 2x_1x_2(3 + x_1 + x_2) - x_1^2x_2^2;$$
$$w_2(x) = x_1 - x_2 + 2x_1x_2(x_2 - x_1).$$

Such polynomials are uniquely determined by their coefficient vectors, which in this case are conveniently portrayed as the 3×3 matrices

$$\begin{array}{c} \\ 1 \\ x_1 \\ x_1^2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & x_2 & x_2^2 \end{array} \\ \begin{pmatrix} 4 & 0 & -1 \\ 0 & 6 & 2 \\ -1 & 2 & -1 \end{pmatrix} \end{array} \quad \text{and} \quad \begin{array}{c} \\ 1 \\ x_1 \\ x_1^2 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & x_2 & x_2^2 \end{array} \\ \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix} \end{array}.$$

Evidently, in this example $\mathcal{W}$ contains two non-overlapping subspaces, one consisting of symmetric functions satisfying $w(x_1, x_2) = w(x_2, x_1)$, the other consisting of skew-symmetric functions satisfying $w(x_1, x_2) = -w(x_2, x_1)$.

0.2.3 *Components of a vector*

Let $\mathcal{V}$ be a vector space of dimension n, and let $\{e_1, \ldots, e_n\}$ be a basis in $\mathcal{V}$. Then each element $v \in \mathcal{V}$ has a unique representation as a linear combination of the basis vectors:

$$v = v^1 e_1 + \cdots + v^n e_n. \tag{1.1}$$

We say that $(v^1, \ldots, v^n)$, or simply v^i, are the components of v with respect to the basis $\{e_1, \ldots, e_n\}$. An essential aspect of tensor notation in this context is that superscripts denote components of a vector in $\mathcal{V}$, and are not to be confused with powers.

It is important at the outset to understand that $(v^1, \ldots, v^n)$, as an ordered set of real numbers in $\mathcal{R}^n$, serves to identify $v \in \mathcal{V}$, but is not in any sense the same point as v because the objects in $\mathcal{V}$ need not be ordered sets of real numbers. The quadratic and bi-quadratic functions at the end of the preceding section illustrate the distinction. The following example is similar but simpler.

Let $\mathcal{V}$ be the space of functions on $[0, 2\pi)$ spanned by the vectors $e_1 = \cos\theta$ and $e_2 = \sin\theta$. Then each element in $\mathcal{V}$ is a linear combination of the form $v(\theta) = 2\cos\theta + 5\sin\theta$, i.e. a function $v(\theta)$ defined on the interval $[0, 2\pi)$. The components of v, in this example $(2, 5)$, may be regarded as a point in $\mathcal{R}^2$, a very different kind of object from a function on the real line. In the bi-quadratic example at the end of the preceding section, the component vector was portrayed as a matrix, quite a different kind of object from a polynomial.

There is another major distinction between the abstract notion of a point $v \in \mathcal{V}$ and its description by means of components $(v^1, \ldots, v^n)$. While it is possible to conceive of a point $v \in \mathcal{V}$ in the abstract, its description via components necessarily involves a basis. The choice of basis is usually arbitrary up to linear transformation. If a new basis is chosen, the description of v by its components must be altered from the original $(v^1, \ldots, v^n)$ to new components $(\bar{v}^1, \ldots, \bar{v}^n)$, say. This transformation is necessary in order that the description $(\bar{v}^1, \ldots, \bar{v}^n)$ with respect to the new basis should identify the same point in $\mathcal{V}$ as $(v^1, \ldots, v^n)$ with respect to the original basis. It is precisely this transformation property that is an automatic and integral part of the tensor calculus.

Despite the conceptual distinction between the vector of components $(v^1, \ldots, v^n)$ and the point $v \in \mathcal{V}$, we shall occasionally refer to 'the point $v = (v^1, \ldots, v^n)$' when in fact we mean 'the point $v \in \mathcal{V}$ whose components with respect to the basis $\{e_1, \ldots, e_n\}$ are $(v^1, \ldots, v^n)$.' Usually the correct meaning is obvious from the context.

0.2.4 *Subspaces and cosets*

We now examine a number or ways in which an existing vector space can be partitioned or extended. Two notions are particularly important in this respect, namely the notion of a vector subspace, and the notion of the span of a collection of vectors.

Definition: A non-empty subset $\mathcal{U}$ of a vector spave $\mathcal{V}$ is a *subspace* if it is closed under linear combinations. In other words, $\mathcal{U}$ is a subspace if, for each pair of points x, y in $\mathcal{U}$, every linear combination $\alpha x + \beta y$ is also in $\mathcal{U}$.

A subspace of a vector space is itself a vector space with the same operations of addition and scalar multiplication derived from the parent space. In particular, every subspace includes the origin. Two extreme subspaces are the set $\mathcal{O}$ consisting of the origin only, and the subspace comprising the whole space $\mathcal{V}$.

The intersection of two or more subspaces is again a subspace, as can be seen from the definition. The union of two subspaces is not ordinarily a subspace because the union does not contain linear combinations. Venn diagrams are thus inappropriate for the representation of subspace operations.

Two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ are said to be *non-overlapping* if the origin is the only point of intersection, i.e. $\mathcal{U}_1 \cap \mathcal{U}_2 = \mathcal{O}$. Note that $\mathcal{O}$ is not the empty set but rather a vector space of dimension zero, so this definition is at odds with the definition of disjoint sets.

DEFINITION: Let U be any subset of $\mathcal{V}$, not necessarily a subspace. The *span* of U, denoted by $\text{span}(U)$, is that subset of $\mathcal{V}$ consisting of linear combinations of the elements of U. If U is a finite set, the span is often written in the form $\text{span}\{u_1, \ldots, u_p\}$.

The span of any collection of vectors in $\mathcal{V}$ is a subspace.

The span of two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ is the set of vectors $x + y$ with $x \in \mathcal{U}_1$ and $y \in \mathcal{U}_2$. This set, denoted by $\mathcal{U}_1 + \mathcal{U}_2$, is a subspace of $\mathcal{V}$. Note that $\mathcal{U} + \mathcal{V} = \mathcal{V}$, and $\mathcal{U} + \mathcal{U} = \mathcal{U}$.

DEFINITION: Let $\mathcal{U}$ be a subspace of $\mathcal{V}$. The set of vectors $x + y$ with $y \in \mathcal{U}$, is called the *translate of* $\mathcal{U}$ *by* x, or a *coset* of $\mathcal{U}$, and is denoted by $x + \mathcal{U}$. If x happens to be an element of $\mathcal{U}$, then $x + \mathcal{U} = \mathcal{U}$, in which case the coset is a vector space. In general, however, if x is not in $\mathcal{U}$, the coset is not a vector space because it does not include the origin. Note that if $x + \mathcal{U} = y + \mathcal{U}$ it does not follow that x and y are the same point in $\mathcal{V}$. It does follow that $x - y$ is a point in $\mathcal{U}$.

For mathematical purposes, the following statistical examples are somewhat simplified to avoid the kinds of complications that invariably arise in applied work.

Example 1: Linear regression. Multiple linear regression refers to a class of statistical models used for studying the dependence of a response variable Y on a (small) number of explanatory variables $x_1, \ldots, x_p$. For present purposes, all of these variables are conceived of as points, vectors or functions in a finite-dimensional vector space $\mathcal{V}$, often but not necessarily equal to R^n. Usually the systematic part of the model is written in the form

$$E(Y \mid x) = \beta_1 x_1 + \cdots + \beta_p x_p$$

where $\beta_1, \ldots, \beta_p$ are unknown scalar coefficients to be estimated. The ultimate aims of multiple regression are many and varied. From our present perspective, if $\beta_1, \ldots, \beta_p$ are unrestricted coefficients, the model states simply that the conditional expected value $E(Y \mid x)$ lies in the subspace defined by the span of the vectors $x_1, \ldots, x_p$. In most cases, x_1 is the constant vector, and β_1 is then called the intercept.

Example 2: Polynomial regression. Let $Y(x)$ be the response observed at the point $x = (x_1, x_2)$ in the plane, or in some subset thereof. For example, Y might denote chemical yield (possibly on a logarithmic scale as a percent of the theoretical maximum), x_1 reaction temperature, and x_2 reaction pressure. Let $\mathcal{U}$ be the space spanned by the functions $1, x_1, x_2, x_1^2, x_2^2, x_1 x_2$. In this case, the response vector, or rather, the expected response vector, is conceived of as a point in the space of functions on 'reaction space', and the subspace $\mathcal{U}$ is the space of polynomials of total degree at most two. In practice, observations are taken at a finite grid of points in the reaction space, in which case $\mathcal{V}$ is the space of functions on this grid, and $\mathcal{U}$ is the subspace of quadratic functions on the same grid.

Example 3: Additive effects model. Let S denote the grid of points $\{1, \ldots, m\} \times \{1, \ldots, n\}$ in the plane, and let $\mathcal{V}$ be the vector space of real-valued functions of two variables on S. Let $\mathcal{U}_1$ be the subspace of functions of one variable x on the points $\{1, \ldots, m\}$, and let $\mathcal{U}_2$ be the subspace of functions of the second variable t on $\{1, \ldots, n\}$. Although we write $f_1(x)$ and $f_2(t)$ for typical elements of $\mathcal{U}_1$ and $\mathcal{U}_2$, it should be borne in mind that f_1 is conceived here as a function on S,

i.e. a function of two variables that happens not to depend on t. Consequently, $\mathcal{U}_0$, the space of functions that are constant on S, is a subspace of both $\mathcal{U}_1$ and $\mathcal{U}_2$. In fact $\mathcal{U}_1 \cap \mathcal{U}_2 = \mathcal{U}_0$. Finally, $\mathcal{U}_1 + \mathcal{U}_2$, the span of the two subspaces, is the space of functions of the form $f_1(x) + f_2(t)$ on S.

In certain areas of applied statistical work, the response at (i, j) is denoted by Y_{ij}, and the additive effects model is written algebraically as $E(Y_{ij}) = \alpha_i + \beta_j$. An alternative, so-called symbolic or model-formula, notation is also employed in the form $A + B$, where A and B are called factors. From the present viewpoint, these factors are vector spaces coinciding with $\mathcal{U}_1$ and $\mathcal{U}_2$, and the operator '+' is the vector span.

0.2.5 *Dimension of a subspace*

We need only cite two facts about subspaces. First, a subspace $\mathcal{U}$ in a vector space $\mathcal{V}$ of dimension n is a vector space of dimension not more than n. Second, given a p-dimensional subspace $\mathcal{U}$ in $\mathcal{V}$, there exists a basis $\{x_1, \ldots, x_p, x_{p+1}, \ldots, x_n\}$ in $\mathcal{V}$ such that $x_1, \ldots, x_p$ are in $\mathcal{U}$ and form a basis for $\mathcal{U}$. (Halmos, 1958, section 12).

0.2.6 *Complementary spaces*

Two non-overlapping vector subspaces having the property that their span coincides with the entire space are said to be *complementary*. In the symbols introduced above, two subspaces $\mathcal{U}_1$ and $\mathcal{U}_2$ are complementary in $\mathcal{V}$ if $\mathcal{U}_1 \cap \mathcal{U}_2 = \mathcal{O}$ and $\mathcal{U}_1 + \mathcal{U}_2 = \mathcal{V}$. If $\mathcal{V}$ is a vector space of dimension n, and $\mathcal{U}_1$ is a subspace of dimension p then the complementary space has dimension $n - p$. Conversely, if $\mathcal{U}_1$ and $\mathcal{U}_2$ are non-overlapping subspaces of dimensions p and $n - p$ in a vector space of dimension n, then $\mathcal{U}_1$ and $\mathcal{U}_2$ are complementary spaces. (Halmos 1958, sections 11, 12).

The importance of complementary subspaces lies in the fact that any vector v in $\mathcal{V}$ can be written as the sum of two components $v = x + y$ with $x \in \mathcal{U}_1$ and $y \in \mathcal{U}_2$. In this respect, both parts of the definition are important. The second part, $\mathcal{U}_1 + \mathcal{U}_2 = \mathcal{V}$ ensures that points $x \in \mathcal{U}_1$ and $y \in U_2$ exist satisfying $x + y = v$ for any $v \in \mathcal{V}$. The non-overlapping condition on the subspaces ensures that the decomposition is unique.

Example 4: Additive effects model continued. In the notation previously established, let $\mathcal{U}_1'$ be the set of functions on $\{1, \ldots, m\}$ satisfying $\sum f_1(x) = 0$. This is a subspace of $\mathcal{U}_1$ satisfying $\mathcal{U}_0 \cap \mathcal{U}_1' = \mathcal{O}$, and $\mathcal{U}_0 + \mathcal{U}_1' = \mathcal{U}_1$, so that $\mathcal{U}_0$ and $\mathcal{U}_1'$ are complementary subspaces of $\mathcal{U}_1$. Likewise, if $\mathcal{U}_2'$ is the set of functions on $\{1, \ldots, n\}$ satisfying $\sum f_2(y) = 0$, $\mathcal{U}_0$ and $\mathcal{U}_2'$ are complementary subspaces of $\mathcal{U}_2$. Also, $\mathcal{U}_0$, $\mathcal{U}_1'$ and $\mathcal{U}_2'$ are complementary subspaces of $\mathcal{U}_1 + \mathcal{U}_2$: they are non-overlapping and their span is $\mathcal{U}_1 + \mathcal{U}_2$. Finally, if we let $\mathcal{U}_{12}'$ be the set of functions on S satisfying

$$\sum_x f(x, y) = 0 \quad \text{for} \quad y = 1, \ldots, n,$$
$$\sum_y f(x, y) = 0 \quad \text{for} \quad x = 1, \ldots, m,$$

it can be seen that $\mathcal{U}_{12}'$ is a subspace complementary to $\mathcal{U}_1 + \mathcal{U}_2$ in $\mathcal{V}$.

As a numerical example with $m = 3$ and $n = 4$ we offer the decomposition

$$\begin{pmatrix} 4 & 2 & 3 \\ 5 & 0 & 1 \\ 3 & 0 & 0 \\ 0 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix} + \begin{pmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 2 & -1 & -1 \\ 1 & 0 & -1 \\ -3 & 1 & 2 \end{pmatrix}$$

The elements on the right, whether they be called functions, vectors or matrices, are in the subspaces $\mathcal{U}_0$, $\mathcal{U}_1'$, $\mathcal{U}_2'$ and $\mathcal{U}_{12}'$ respectively.

Back now to the general case, if $\mathcal{U} = \mathcal{V}$ there is no complementary subspace other than the trivial subspace $\mathcal{O}$ consisting of the origin only. Likewise, if $\mathcal{U} = \mathcal{O}$, the only complementary subspace is $\mathcal{V}$ itself. In general, however, for a given subspace $\mathcal{U}$ of a vector space $\mathcal{V}$, there exist infinitely many subspaces that are complementary to $\mathcal{U}$ in $\mathcal{V}$. For example, if $\mathcal{V}$ is the real plane and $\mathcal{U}$ is the real axis, any line through the origin other than the horizontal axis can serve as a complement of $\mathcal{U}$. If we take the y-axis as a complement to $\mathcal{U}$, a point (x, y) in $\mathcal{V}$ has the familiar decomposition $(x, 0) + (0, y)$ as the sum of a vector in $\mathcal{U}$ and a vector in the complement. If, on the other hand, we choose the equi-angular line (x, x) as the complement to $\mathcal{U}$, the decomposition becomes

$$(x, y) = (x - y, 0) + (y, y).$$

In the additive effects example described above, the spaces $\mathcal{U}_1$, $\mathcal{U}_2$ and their span are prescribed by subject-matter coinsiderations. But the complementary space $\mathcal{U}_{12}'$ depends on a simple, but ultimately arbitrary, choice. A different choice of complementary space would yield a different decomposition.

0.2.7 *Quotient space*

Let $\mathcal{V}$ be a vector space. For a given subspace $\mathcal{U}$ there are, usually, infinitely many complementary subspaces. Without some additional structure, there is no natural way to select one of these complements for preferential treatment. There is, however, a natural way of patching together all these complementary subspaces to form a new vector space that, for all practical purposes, gives a unique complement of $\mathcal{U}$ in $\mathcal{V}$.

The idea is to manufacture a new vector space whose elements are not *points* in $\mathcal{V}$, but rather *subsets* of $\mathcal{V}$, in fact the cosets of $\mathcal{U}$. To do so, we need to define the vector space operations on cosets of $\mathcal{U}$. Addition of two cosets is defined by addition of the elements: this operation yields a new coset. Scalar multiplicaton is defined likewise, again yielding a new coset. Thus

$$(v_1 + \mathcal{U}) + (v_2 + \mathcal{U}) = v_1 + v_2 + \mathcal{U} + \mathcal{U} = v_2 + v_2 + \mathcal{U},$$
$$\alpha(v_1 + \mathcal{U}) = \alpha v_1 + \alpha\mathcal{U} = \alpha v_1 + \mathcal{U}$$

for any scalar α. The cosets thus form a vector space, $\mathcal{U}$ itself being the zero element. This vector space is called the *quotient space*, and is denoted by the symbol $\mathcal{V}/\mathcal{U}$. The dimension of the quotient space is $\dim(\mathcal{V}) - \dim(\mathcal{U})$, the same as the dimension of any complementary space. (Halmos, 1958, sections 21, 22).

One way to visualize the quotient space, illustrated graphically in Fig 1.2, is as follows. Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be two vector spaces, both complementary to $\mathcal{U}$ in $\mathcal{V}$. Since $\mathcal{V} = \mathcal{U} + \mathcal{W}_1 = \mathcal{U} + \mathcal{W}_2$, any point v in $\mathcal{V}$ can be expressed as the sum of a pair of points in two ways:

$$v = u_1 + w_1 = u_2 + w_2,$$

where u_1, u_2 are in $\mathcal{U}$, $w_1 \in \mathcal{W}_1$, and $w_2 \in \mathcal{W}_2$. It follows then that $w_1 - w_2 = u_1 - u_2$ is a point in $\mathcal{U}$. Consequently $w_1 + \mathcal{U}$ and $w_2 + \mathcal{U}$ are two representations of the same coset, and $v + \mathcal{U}$ is a third representation of the same coset. The coset associated with a point v is thus the set of projections of $\mathcal{V}$ along $\mathcal{U}$ on to the various complementary spaces. The cosets of $\mathcal{U}$ partition the space into equivalence classes: two points in $\mathcal{V}$ are equivalent in this sense if their difference is in $\mathcal{U}$.

Example 5: Linear regression with offset. Consider the linear regression model

$$E(Y \mid x) = x_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where $x_0, \ldots, x_p$ are vectors in $\mathcal{V}$. If $\mathcal{U} = \text{span}\{x_1, \ldots, x_p\}$, this model asserts that the mean of Y lies on a particular coset of $\mathcal{U}$. Since x_0 has unit coeficient, the model specifies the coset $x_0 + \mathcal{U}$, but it does not otherwise identify the point on the coset.

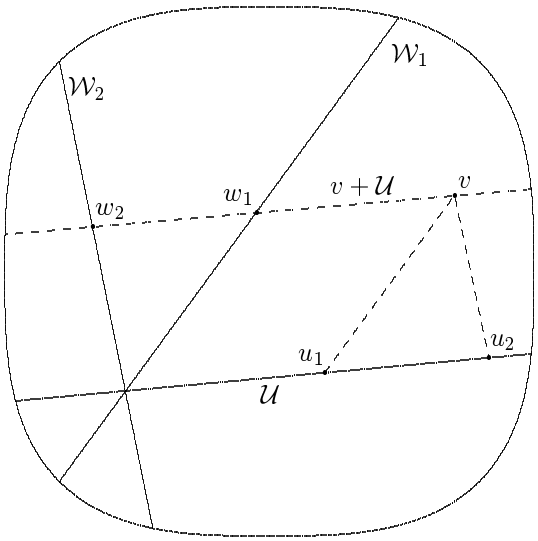

Figure 1.1. *Stone diagram illustrating a subspace $\mathcal{U}$ and two complementary spaces $\mathcal{W}_1$ and $\mathcal{W}_2$. The coset $v+\mathcal{U}$ is illustrated by a dashed line parallel to $\mathcal{U}$. The two decompositions $v = u_1 + w_1$ and $v = u_2 + w_2$ are also illustrated geometrically by the parallelogram rule for vector addition.*

Example 6: Residual vector in linear models. Consider the linear model, written in conventional matrix notation as

$$Y = X\beta + \epsilon,$$

where X is a given matrix whose columns span a subspace $\mathcal{X}$ of $\mathcal{V} = \mathcal{R}^n$, and ϵ is a random variable in $\mathcal{R}^n$, whose distribution depends on unknown parameters, θ. Let P be any projection matrix having range $\mathcal{X}$, and let $Q = I - P$ be the complementary, or residual, projection having null space $\mathcal{X}$. For example, we could take $P = X(X^TX)^{-1}X^T$ or $P' = X(X^TWX)X^TW$ for any symmetric positive definite matrix W. Whatever the choice of P, $R = QY$ is a residual vector. Since $R = Q\epsilon$, the residual vector has a distribution independent of β, but dependent on θ. Unlike the vector of fitted values whose definition depends on the choice of inner product or weight matrix W, all residual vectors are equivalent in the sense that they contain the same information. For example, the residual $R' = Q'Y$ can be obtained from $R = QY$ by the linear transformation $R' = Q'R$, and conversely $R = QR'$. Note that, although R' and R are vectors in different spaces, the difference $R' - R$ is necessarily a linear combination of the columns of X, i.e. an element of $\mathcal{X}$. For this reason, it is generally preferable to regard the residual vector as an element of the quotient space $\mathcal{V}/\mathcal{X}$ rather than an element of any particular subspace complementary to $\mathcal{X}$ in $\mathcal{V}$. By this device, we make precise the notion that all residual vectors are equivalent.

From a slightly different point of view, a residual projection on $\mathcal{V}$ having null space $\mathcal{X}$ transforms all points in a given coset to the single point where the range intersects the coset. Since all points in a coset are transformed to the same point in the range, the transformation on $\mathcal{V}$ can alternatively be considered as a transformation from the quotient space to itself. From this point of view, each residual projections on $\mathcal{V}$ acts as the identity on $\mathcal{V}/\mathcal{X}$, so all residual projections on $\mathcal{V}$ are identical on $\mathcal{V}/\mathcal{X}$.

In fact, it is technically not necessary to construct a projection matrix or form a residual vector at all: we may talk simply of the distribution of the random variable $Y + \mathcal{X}$ in $\mathcal{V}/\mathcal{X}$, and use the resulting 'residual likelihood' for purposes of estimating θ. So far as its dependence on the data is concerned, the residual likelihood is a function on the space $\mathcal{V}/\mathcal{X}$, i.e. constant on the cosets of $\mathcal{X}$. We return to this example in section ?.?.

0.2.8 *Direct sum and tensor sum of vector spaces*

Given two vector spaces, we may form a new, and larger, vector space, in at least two different ways. The first of these sums is called the direct sum or Cartesian product, and is defined as follows.

DEFINITION: *Direct sum of vector spaces.* Let $\mathcal{U}$ and $\mathcal{V}$ be two vector spaces over the same field of scalars. The direct sum $\mathcal{W} = \mathcal{U} \oplus \mathcal{V}$ is a new vector space whose elements are all the ordered pairs (u, v) with $u \in \mathcal{U}$ and $v \in \mathcal{V}$. The linear operations are defined component-wise

$$\alpha(u_1, v_1) + \beta(u_2, v_2) = (\alpha u_1 + \beta u_2,\ \alpha v_1 + \beta v_2)$$

in the obvious way.

The subspaces consisting of the elements $(u, 0)$ and $(0, v)$ may be identified with $\mathcal{U}$ and $\mathcal{V}$ respectively. In this sense, although there is a slight technical innacuracy, we may speak of $\mathcal{U}$ and $\mathcal{V}$ as non-overlapping subspaces of $\mathcal{W}$, and $\mathcal{W} = \mathcal{U} + \mathcal{V}$, the span of the two complementary subspaces. The dimension of $\mathcal{U} \oplus \mathcal{V}$ is $\dim(\mathcal{U}) + \dim(\mathcal{V})$.

The Cartesian product of two sets U and V, in general not vector spaces, is the set of ordered pairs (u, v) with $u \in U$ and $v \in V$. So far as it goes, this coincides with the definition of direct sum. Unlike the direct sum of vector spaces, however, the term Cartesian product implies no linear operations.

It is worth pausing for a moment to examine the nature of the elements of the direct sum space just defined. The definition of direct sum is exactly what is required for constructing the real plane from the real line, or, by extension, $\mathcal{R}^n$ from n copies of the real line. Suppose, however, that the elements of $\mathcal{U}$ and $\mathcal{V}$ are real-valued functions on domains D_1 and D_2. According to the definition, the elements of $\mathcal{U} \oplus \mathcal{V}$ are ordered pairs of functions, in other words, functions from $D_1 \times D_2$ taking values in the plane. In a real sense, the elements of $\mathcal{U} \oplus \mathcal{V}$ are quite unlike the elements of the component spaces. In addition, even though the elements of $\mathcal{U}$ and $\mathcal{V}$ both take values in the reals, elements of the complementary spaces $(\mathcal{U}, 0)$ and $(0, \mathcal{V})$ do not take values in the same set. These considerations lead to a different kind of sum, what we shall call the tensor sum of two vector spaces.

Let 1 be the one-dimensional vector space of functions that are constant on D_2. Then $\mathcal{U} \otimes 1$ is the vector space of functions on $D_1 \times D_2$ of the form $f(x_1, x_2) = g(x_1)$ for $g \in \mathcal{U}$. Likewise, $1 \otimes \mathcal{V}$ is the vector space of functions on $D_1 \times D_2$ of the form $f(x_1, x_2) = h(x_2)$ for $h \in \mathcal{V}$. The span $\mathcal{U} \otimes 1 + 1 \otimes \mathcal{V}$, more commonly denoted in statistical work by $\mathcal{U} + \mathcal{V}$, is the vector space of functions on $D_1 \times D_2$ of the form $g(x_1) + h(x_2)$.

DEFINITION: *Tensor sum of two function spaces.* Let $\mathcal{U}$ and $\mathcal{V}$ be vector spaces of real-valued functions on domains D_1 and D_2 respectively. The tensor sum $\mathcal{U} + \mathcal{V}$ is the vector space of functions of the form $u + v$ with $u \in \mathcal{U}$ and $v \in \mathcal{V}$, acting on the domain $D_1 \times D_2$.

With the domain extended in this way, the tensor sum is simply the span of the two subspaces, so there is no need for a new symbol for this operation. The important condition is that the vector spaces $\mathcal{U}$ and $\mathcal{V}$ should be such that addition of elements from the two spaces is defined in a way that is compatible with addition in each of the component spaces. In other words, $\mathcal{U}$ and $\mathcal{V}$ must be vector spaces of functions. Otherwise the tensor sum is not defined.

The example that follows shows that, in general, $\mathcal{U}$ and $\mathcal{V}$ have a non-trivial intersection, so the dimension of the tensor sum is less that the sum of the dimensions of the component spaces. The dimension of the tensor sum is

$$\dim(\mathcal{U} + \mathcal{V}) = \dim(\mathcal{U}) + \dim(\mathcal{V}) - \dim(\mathcal{U} \cap \mathcal{V}).$$

Example 7: Direct sum versus tensor sum. Let $\mathcal{U}$ be the span of the polynomials $\{1, x, x^2\}$ on $[0, 1]$, and let $\mathcal{V}$ be the span of $\{1, y\}$ on $\mathcal{R}$. By the definition given above, the elements of $\mathcal{U} \oplus \mathcal{V}$

are ordered pairs of functions $(p_2(x), p_1(y))$ from the set $[0,1] \times \mathcal{R}$ into $\mathcal{R}^2$. Here, p_r denotes a polynomial of degree not exceeding r. A typical element of the direct sum $\mathcal{U} \oplus \mathcal{V}$ has the form

$$(a_0 + a_1 x + a_2 x^2,\ b_0 + b_1 y)$$

with arbitrary coefficients a_0, a_1, a_2, b_0, b_1. The space clearly has dimension 5, the sum of the dimensions of the component spaces.

The tensor sum, on the other hand, is the set of functions $p_2(x) + p_1(y)$ from $[0,1] \times \mathcal{R}$ into $\mathcal{R}$. This space is spanned by the functions $\{1, x, x^2, y\}$, and is thus of dimension 4. Note that functions in $\mathcal{U}$ are constant in y, and functions in $\mathcal{V}$ are constant in x. The intersection, $\mathcal{U} \cap \mathcal{V}$, the space of functions that are constant in both variables, has dimension 1.

Example 8: Randomized blocks model:. In a randomized blocks design, the expected response when treatment j is applied to block i is often modelled as an additive function of block and treatment effects. To put this in the language of vector spaces, let $\mathcal{U}$ be the space of functions on the block labels, and let $\mathcal{V}$ be the space of functions on the treatments. Then $\mathcal{U} + \mathcal{V}$ is the space of functions $u(i) + v(j)$ on the Cartesian product of the domains, $\{1, \ldots, m\} \times \{1, \ldots, n\}$. In the notation commonly used in statistics for specifying linear models, this space is denoted by $B + T$ where B is the factor for blocks and T is the factor for treatment. A factor in this context is the vector space of functions on a finite set of labels, and the operator '+' denotes the span or tensor sum.

0.2.9 *Dual basis*

The following theorem is taken from Halmos (1958, section 15).

THEOREM 1. Let $\mathcal{V}$ be an n-dimensional vector space and let $\{e_1, \ldots, e_n\}$ be a basis in $\mathcal{V}$. Then there is a uniquely determined basis $\{y^1, \ldots, y^n\}$ in $\mathcal{V}^*$ with the property that $y^i(e_j) = \delta^i_j$. ($\delta^i_j = 1$ for $i = j$ and zero otherwise is an extremely useful symbol known as Kronecker's delta.) Consequently the dual space is n-dimensional.

For a proof of this assertion, see Halmos (1958, section 15).

The basis $\{y^1, \ldots, y^n\}$, which is the set of coordinate projections on to $\{e_1, \ldots, e_n\}$, is called the *dual basis* in $\mathcal{V}^*$ of $\{e_1, \ldots, e_n\}$. If $\mathcal{V}$ is a real inner product space with inner product $\langle \cdot, \cdot \rangle$, and if G is the matrix with elements $G_{ij} = \langle e_i, e_j \rangle$, then the coordinate projections are given by

$$y^i(v) = \sum_j G^{ij} \langle e_j, v \rangle, \tag{1.2}$$

where G^{ij} is the (i, j) element of the matrix G^{-1}.

0.2.10 *Examples*

The following examples are of a non-statistical nature and are intended to clarify the distinction between $\mathcal{V}$ and $\mathcal{V}^*$.

Example 9: Harmonics on the circle. Let $\mathcal{V}$ be the space of zero and first-order harmonics on the unit circle. In symbols,

$$\mathcal{V} = \text{span}\{e_0 = 1, e_1 = \cos\theta, e_2 = \sin\theta\}$$

for $0 \leq \theta < 2\pi$. We can think of the elements of $\mathcal{V}$ either as periodic functions defined on an interval of the real line, or preferably, as continuous functions on the unit circle. Typical elements of $\mathcal{V}$ include functions such as $\sin(\theta + \pi/4)$ and $\cos^2(\theta/2)$. Every vector in $v \in \mathcal{V}$ can be expressed in a unique way as a linear combination of the basis vectors

$$v(\theta) = v^0 1 + v^1 \cos\theta + v^2 \sin\theta.$$

The following are examples of linear functionals on $\mathcal{V}$.

$$\begin{aligned} y^0(v) &= \frac{1}{2\pi}\int_0^{2\pi} v(\theta)\,d\theta, \\ y^1(v) &= \frac{1}{\pi}\int_0^{2\pi} v(\theta)\cos\theta\,d\theta, \\ y^2(v) &= \frac{1}{\pi}\int_0^{2\pi} v(\theta)\sin\theta\,d\theta, \\ \int_0^{2\pi} v(\theta)\cos^2(\theta/2)\,d\theta &= \pi y^0(v) + \pi y^1(v)/2. \end{aligned}$$

In terms of the basis vectors $\{e_0, e_1, e_2\}$, in $\mathcal{V}$, the value of the ith linear functional on the jth basis vector is $y^i(e_j) = \delta^i_j$. Thus $\{y^0, y^1, y^2\}$ are the dual basis vectors in $\mathcal{V}^*$. Consequently, $y^j(v) = v^j$, the jth component of v with respect to the given basis vectors. For instance, the fourth linear functional in the preceding list has components $(\pi, \pi/2, 0)$ with respect to the dual basis. In other words, an element v^* in $\mathcal{V}^*$ with coefficient vector (v^*_0, v^*_1, v^*_2) is a linear functional

$$\begin{aligned} v^*(v) &= v^*_0 y^0(v) + v^*_1 y^1(v) + v^*_2 y^2(v) = v^*_0 v^0 + v^*_1 v^1 + v^*_2 v^2 \\ &= \frac{1}{2\pi}\int_0^{2\pi} (v^*_0 + 2v^*_1\cos\theta + 2v^*_2\sin\theta) v(\theta)\,d\theta. \end{aligned}$$

It is common practice to use the word 'coefficient' for the components of a vector in the dual space. Each coefficient vector represents a different linear combination or linear functional.

Example 10: Contrast vectors. Let $\mathcal{V}$ be the set of contrast vectors in $\mathcal{R}^3$. An element of $\mathcal{V}$ is an ordered set of real numbers $(\bar{v}^1, \bar{v}^2, \bar{v}^3)$ whose sum is zero. We take as basis vectors the two linearly independent contrast vectors

$$e_1 = (1, -1, 0) \quad \text{and} \quad e_2 = (-1, 0, 1).$$

Any contrast vector $v = (\bar{v}^1, \bar{v}^2, \bar{v}^3)$ can be written as a linear combination of e_1 and e_2:

$$v = v^1 e_1 + v^2 e_2.$$

Using (1.2), the coefficients, which are linear functionals on $\mathcal{V}$, are given by

$$\begin{aligned} v^1(v) &= (\bar{v}^1 - 2\bar{v}^2 + \bar{v}^3)/3, \\ v^2(v) &= (-\bar{v}^1 - \bar{v}^2 + 2\bar{v}^3)/3. \end{aligned}$$

Since $v^i(e_j) = \delta^i_j$, $\{v^1, v^2\}$ is the dual basis in $\mathcal{V}^*$ of $\{e_1, e_2\}$. In other words, the point $\alpha = (\alpha_1, \alpha_2)$ in $\mathcal{V}^*$ denotes the linear functional

$$\alpha(v) = \alpha_1 v^1(v) + \alpha_2 v^2(v).$$

0.2.11 *Notation and terminology*

In statistical contexts, variables have names such as 'degrees Celsius,' 'kilometres east,' or 'log weight in kg.' These names form a basis in $\mathcal{V}$ called the incidence basis, denoted here by $\{e_1, \ldots, e_p\}$, where p is the number of variables. Thus, e_i is synonymous with the term 'name of variable i', and v^i is synonymous with 'value of variable i,' or 'value of v on variable i.' The complete description of $v \in \mathcal{V}$ is thus $v = v^i e_i$, or equivalently, $(v^1, \ldots, v^p)$, the basis vectors being understood from the context. Stone (1987) uses the notation v_i in place of e_i for the name of variable i. By a slight modification of notation in Halmos (1958), $[e, v_i]$ is used by Stone in place of our v^i for the value on variable i.

According to theorem 2, every linear functional on $\mathcal{V}$ can be written as a linear combination of the dual basis vectors or coordinate projections

$$\alpha(v) = \alpha_1 v^1 + \cdots + \alpha_n v^n = \alpha_i v^i,$$

with implicit summation over the repeated index. Each element in the dual space may therefore be identified with a coefficient vector $\alpha = (\alpha_1, \ldots, \alpha_n)$ in $\mathcal{R}^n$. These are the components of the linear functional $\alpha \in \mathcal{V}^*$ with respect to the dual basis $\{v^1, \ldots, v^n\}$. Here, we have identified the dual space with the set of coefficient vectors α. Stone (1987) identifies $\mathcal{V}^*$ with the space of 'evaluators,' so that our linear functional $\alpha_i v^i$ coincides with his $[\alpha, v]$.

In tensor notation, the components of $v \in \mathcal{V}$ are identified by means of superscripts. The summation convention then permits (1.1) to be abbreviated to $v = v^i e_i$. The components of a linear functional $\alpha \in \mathcal{V}^*$ are denoted by α_i, a coefficient vector with subscripts. The linear functional $\alpha(v)$ may therefore be written in the form $\alpha(v) = \alpha_i v^i$. To achieve uniformity of notation and to blend in with the summation convention, components of vectors in $\mathcal{V}$ are indexed by superscripts: components of vectors in $\mathcal{V}^*$, also called coefficients, are indexed by subscripts. The distinction made in matrix notation between row vectors and column vectors is related to, but does not coincide uniformly with the distinction between component vectors in $\mathcal{V}$ and $\mathcal{V}^*$. For further discussion on this point, see section 1.4.4.

It may be helpful at this point to digress briefly on the subject of matrix notation. Consider for example, the identity matrix of order n, written in matrix notation as I or I_n. Such a matrix arising as a linear transformation from $\mathcal{R}^n$ to $\mathcal{R}^n$ is written in index notation using Kronecker's delta in the form δ^i_j. However, I_n could also arise as a covariance matrix of the random variables $(X^1, \ldots, X^n)$, in which case we would write $\delta^{ij} = \mathrm{cov}(X^i, X^j)$ with superscripts since all tensor equations must be balanced with respect to the positions of indices. Although δ^i_j is numerically equal to δ^{ij}, the two symbols are not equal as tensors because it is implicit in the notation that they are designed to operate in quite different ways. To say the same thing in a different way, tensor notation implies or suggests that these are the components of vectors in different vector spaces. The fact that the components are numerically equal is then irrelevant when the vectors in question do not lie in the same space.

0.3 Transformation rules

0.3.1 *Duality and dual spaces*

Let $\mathcal{V}$ be a vector space with basis $\{e_1, \ldots, e_n\}$. In the usual expression for $v \in \mathcal{V}$ as a linear combination of the basis vectors, we write $v = v^i e_i$. The notation exhibits a certain formal symmetry between the components v^i and the basis vectors e_i, but the interpretation, as usually given, is quite asymmetric: v^i is a 'scalar' whereas e_i is a 'vector' in $\mathcal{V}$. However, $\{v^1, \ldots, v^n\}$, as linear functionals, form a basis in $\mathcal{V}^*$, in fact the dual basis of $\{e_1, \ldots, e_n\}$. Thus, the expression $v^i e_i$ could equally well be interpreted as a point in $\mathcal{V}^*$ with components $(e_1, \ldots, e_n)$ relative to the dual basis. Tensor notation permits, and even encourages, this dual interpretation. In this dual

sense, $v^i e_i$ is formally an invariant. It identifies a point in $\mathcal{V}$ independent of the choice of basis: it also identifies a point in $\mathcal{V}^*$ independent of the basis.

The following arguments, although they do not attempt to exploit the symmetry of the duality described above, explain why expressions such as $v^i e_i$, having no free index, are invariant.

0.3.2 *Change of basis vectors in* $\mathcal{V}$

Let $\mathcal{V}$ be a vector space of dimension n. The general idea is that we want to talk of a point in $\mathcal{V}$ in a way that is automatically independent of the choice of basis vectors. At the same time the methods employed must be sufficiently concrete to permit computation in the usual ways. Obviously the *representation* of $v \in \mathcal{V}$ in terms of its components $(v^1, \ldots, v^n)$ with respect to $\{e_1, \ldots, e_n\}$ does depend on the choice of basis vectors: if the basis vectors are altered, the *representation* will also be altered, but the point itself remains unchanged.

Introduce a new set of basis vectors $\{\bar{e}_1, \ldots, \bar{e}_n\}$. Each member of the new basis is a linear combination of the old basis with coefficients K as follows:

$$\bar{e}_r = K^j_r e_j. \tag{1.3}$$

The inverse transform giving the old basis vectors in terms of the new is also linear, but with coefficients L:

$$e_j = L^r_j \bar{e}_r.$$

On substituting one of these into the other we obtain

$$\bar{e}_r = K^j_r L^s_j \bar{e}_s, \quad \text{and} \quad e_j = K^i_r L^r_j e_i.$$

In other words K and L are matrix inverses. We write $K^j_r L^s_j = \delta^s_r$, $K^i_r L^r_j = \delta^i_j$, where δ^i_j is the identity matrix, or Kronecker's delta. In contrast to matrix notation, the order of symbols in a tensor formula is irrelevant because scalar multiplication is commutative.

Let $(v^1, \ldots, v^n)$ be the components of v with respect to the basis $\{e_1, \ldots, e_n\}$, and let $(\bar{v}^1, \ldots, \bar{v}^n)$ be the components of the same point with respect to $\{\bar{e}_1, \ldots, \bar{e}_n\}$. Then we have

$$v = v^i e_i = v^i L^r_i \bar{e}_r = \bar{v}^r \bar{e}_r.$$

In other words, under the change of basis (1.3) with matrix K, the components of v are transformed linearly from v^i to $\bar{v}^i$ by

$$\bar{v}^r = L^r_i v^i. \tag{1.4}$$

This linear transformation involves the matrix *inverse* of K in (1.3). It is this transformation that is implied when we say that $(v^1, \ldots, v^n)$ is a *contravariant vector* or *contravariant tensor* of order one.

The dual space is associated with coefficient vectors whose components are $(\alpha_1, \ldots, \alpha_n)$ with respect to the dual basis in $\mathcal{V}^*$ of $\{e_1, \ldots, e_n\}$. When the basis vectors in $\mathcal{V}$ are transformed by (1.3) there is a corresponding change in the dual basis vectors from v^i to $\bar{v}^i$ as in (1.4) or (1.2). By the argument given above, the components $(\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$ with respect to the new basis vectors are given by

$$\bar{\alpha}_r = K^j_r \alpha_j, \tag{1.5}$$

involving the matrix inverse of L in (1.4). The net effect of these two changes is that the value of any linear functional such as $\alpha_i v^i$ is unchanged:

$$\alpha_i v^i = \alpha_i K^i_r \bar{v}^r = \bar{\alpha}_i \bar{v}^i.$$

It is transformation (1.5) that is implied when we say that the coefficient vector $(\alpha_1, \ldots, \alpha_n)$, or simply α_i, is a *covariant vector* or *covariant tensor* of order one.

A tensor is thus a component vector used to identify points in a vector space or in the dual space. Implicit in the tensorial recipe is the notion that the basis vectors are arbitrary up to linear transformation. In order that the description v^i should represent the same point in $\mathcal{V}$ when a new basis is selected according to (1.3), it is necessary for the description to change linearly by rule (1.4). Likewise, in the dual space of coefficient vectors, in order that the vector α should represent a specific linear functional independent of the choice of basis, it is necessary that the components α_i undergo transformation by rule (1.5) when a new basis is introduced in $\mathcal{V}$. These transformation rules are implicit in tensor notation. One advantage of the notation is that it permits us to use the basis vectors for computational purposes. At the same time, the explicit recognition that the basis vectors are arbitrary up to linear transformation effectively means that all calculations are independent of the choice of basis. In particular, all scalars are invariant, i.e. independent of the choice of basis.

0.3.3 *Group action*

An alternative description, leading essentially to the same result, can be phrased in terms of group action. Let $\mathcal{G}$ be a group acting on a real n-dimensional vector space $\mathcal{V}$. That is to say that for each $g \in \mathcal{G}$ and $v \in \mathcal{V}$, gv is a point in $\mathcal{V}$. Further, $g_2(g_1 v)$ is the same point as $(g_2 g_1)v$. The elements of the group often have a physical description or interpretation such as 'transposition,' 'cyclic permutation,' 'rotation,' 'reflection,' 'dilation,' 'isotropic expansion' and so on. Such a group often, but not always, has a representation in terms of $n \times n$ matrices acting on $\mathcal{R}^n$ in the sense of matrix multiplication. It is an essential part of any representation that the matrix representation of the group product $g = g_2 g_1$ be the same as the matrix product of the representations of g_2 and g_1. Assuming that $\mathcal{G}$ has such a matrix representation, we may think of g as a matrix, v as a column vector in $\mathcal{V} = \mathcal{R}^n$, and gv representing matrix multiplication. The action of g is denoted by the symbol $v \mapsto gv$ for each $v \in \mathcal{V}$.

Any action of the group on $\mathcal{V}$ is accompanied by an 'equal and opposite reaction' on $\mathcal{V}^*$. In the space of linear functionals on $\mathcal{V}$, the components of the linear functional α with respect to the dual basis are denoted by $(\alpha_1, \ldots, \alpha_n)$, i.e. a vector in $\mathcal{R}^n$. It is convenient in our matrix representation to take α to be a row vector. This helps to distinguish $\mathcal{V}$ from $\mathcal{V}^*$, both being isomorphic to $\mathcal{R}^n$. The actions of g on $\mathcal{V}$ and $\mathcal{V}^*$ are given by

$$g : v \mapsto gv; \quad \alpha \mapsto \alpha g^{-1}.$$

In tensor notation with superscripts indexing rows and subscripts indexing columns,

$$g : v^i \mapsto g^i_r v^r; \quad \alpha_i \mapsto h^r_i \alpha_r, \tag{1.6}$$

where g^i_r are the elements of the matrix g and h^r_i are the elements of g^{-1}.

In this representation, a linear functional on $\mathcal{V}$ is given by the product αv, or in tensor notation by $\alpha_i v^i$. It is immediately apparent from the definition of group action that linear functionals are invariant under the group. With respect to the action of the group $\mathcal{G}$, component vectors in $\mathcal{V}$ are said to be contravariant; component vectors in $\mathcal{V}^*$ are said to be covariant.

In this description there is no suggestion of a change of basis, so v and gv represent different points in $\mathcal{V}$. The emphasis is thus quite different from that in the previous section where v^i and $\bar{v}^i$ were taken to be alternative representations of the same point in $\mathcal{V}$, but with respect to different bases. The primary emphasis in both cases is on scalars, i.e. invariants either under the action of $\mathcal{G}$, or under change of basis. If $\mathcal{G}$ is the group of all linear transformations on $\mathcal{R}^n$, the matrices g and h in (1.6) may be identified with L and K in (1.3) and (1.5) respectively. In this context at least, group action and change of basis are alternative ways of motivating the tensor transformation rules (1.6), or (1.4), (1.5). Although the motivations are very different, the aims are identical and the net result, the tensor transformation rules, is the same in both cases.

The present description in terms of group action is a little more flexible, however, because we may tailor the group to suit the circumstances. In some physical contexts the orthogonal group may be appropriate, particularly where 'real space' $\mathcal{R}^3$ is involved. The matrices appearing in (1.6) are then required to be orthogonal, so that scalars are invariant under rotation and reflection, but not necessarily under dilation or linear transformation. Where moments and cumulants of random variables are involved, the affine group of transformations is usually the appropriate choice for $\mathcal{G}$. In other contexts, the group of lower triangular matrices may be the appropriate choice.

0.3.4 *Adjoint transformation*

Let A be a linear transformation on a vector space $\mathcal{V}$, i.e. $A\colon \mathcal{V} \mapsto \mathcal{V}$. For any linear functional α, there exists an associated linear functional $A'\alpha$ such that the numerical value of α at Ax is equal to the value of $A'\alpha$ at x. Further, A' is a linear transformation on $\mathcal{V}^*$, called the adjoint linear transformation of A (Halmos, 1958, section 44).

Different notational conventions express this notion in different ways. Halmos writes $[x, \alpha]$ to denote the value of the linear functional α at x where others might write $\alpha(x)$. Then the definition is $[Ax, \alpha] = [x, A'\alpha]$, and the assertion is that A' is a linear transformation on $\mathcal{V}^*$.

In one form of matrix notation, elements of $\mathcal{V}$ are column vectors and elements of $\mathcal{V}^*$ are row vectors. The linear transformation A on $\mathcal{V}$ is represented by matrix multiplication Ax, and $[x, \alpha]$ is the scalar αx. Then the value of α at Ax is αAx, which is the same as αA evaluated at x. The adjoint linear transformation $A'\alpha$ on $\mathcal{V}^*$ is thus αA, or matrix post-multiplication.

Using index notation, the components of Ax are $a^i_r x^r$ and the value of α at x is $\alpha_i x^i$. Then $[Ax, \alpha] = \alpha_i a^i_r x^r = [x, A'\alpha]$. The adjoint operation $A'\alpha$ on $\mathcal{V}^*$ is thus $a^i_r \alpha_i$. So the adjoint operation is an immediate and automatic consequence of this notational convention. No transposition of indices is involved.

0.3.5 *Examples*

The examples that follow are intended to illustrate the variety of objects that can be considered to be tensors.

Example 11: Vector product in $\mathcal{R}^3$. Let $\mathcal{V} = \mathcal{R}^3$ represent real space in the Newtonian sense. For any two elements $a = (a^1, a^2, a^3)$ and $b = (b^1, b^2, b^3)$ in $\mathcal{V}$, define the vector product or cross product $c = a \times b$ by its components

$$\begin{aligned} c^1 &= a^2 b^3 - a^3 b^2, \\ c^2 &= a^3 b^1 - a^1 b^3, \\ c^3 &= a^1 b^2 - a^2 b^1. \end{aligned}$$

Certainly c is an ordered triple of real numbers. By some definitions, that property alone would qualify c for membership of $\mathcal{V}$, but our requirements are more stringent than this. Use of superscripts suggests that c is a contravariant vector in $\mathcal{V}$, but this is far from obvious from the definition. The question at issue, then, is whether or not c identifies the same point in $\mathcal{V}$ when we select a new basis for $\mathcal{V}$.

Consider changing basis vectors from e_r to $\bar{e}_r$ as in (1.3). Components with respect to this new basis are denoted by $\bar{a}^i$, $\bar{b}^i$ and $\bar{c}^i$. Since a and b are the components of vectors in $\mathcal{V}$, they transform by rule (1.4) for contravariant vectors. Hence, from the definition of the cross product, the components of c with respect to the new basis are

$$\begin{aligned} \bar{c}^1 &= \bar{a}^2 \bar{b}^3 - \bar{a}^3 \bar{b}^2 \\ \bar{c}^2 &= \bar{a}^3 \bar{b}^1 - \bar{a}^1 \bar{b}^3, \\ \bar{c}^3 &= \bar{a}^1 \bar{b}^2 - \bar{a}^2 \bar{b}^1. \end{aligned}$$

Simplification of the expression for $\bar{c}^1$ gives

$$\bar{c}^1 = L_i^2 a^i L_j^3 b^j - L_j^3 a^j L_i^2 b^i = L_i^2 L_j^3 (a^i b^j - a^j b^i),$$

showing that $\bar{c}^1$ is a linear function of $\{c^1, c^2, c^3\}$. In order for c^i to be a tensor or vector in $\mathcal{V}$ its components must transform to $\bar{c}^i = L_r^i c^r$. The above expression, which is quadratic in L, shows that this is not the case for general L, so the cross product is not a contravariant vector in the sense of section 1.3.

A question of interest related to the discussion of group action is the following. For what group of transformations, or set of matrices L, does c obey the transformation law of a contravariant vector? We now investigate this point in detail.

Further simplification of the expression for $\bar{c}^1$ gives

$$\bar{c}^1 = (L_1^2 L_2^3 - L_2^2 L_1^3) c^3 + (L_3^2 L_1^3 - L_1^2 L_3^3) c^2 + (L_2^2 L_3^3 - L_3^2 L_2^3) c^1.$$

There are similar expressions for $\bar{c}^2$ and $\bar{c}^3$. Now, if L is 3×3 and orthogonal, it can be shown that L_j^i is equal to its co-factor multiplied by $|L|$. Thus, if L is orthogonal, we have

$$\bar{c}^i = |L| \, L_r^i c^r.$$

In other words, $c = a \times b$ is a contravariant vector in $\mathcal{V}$, but only with respect to O^+, the group of rotations in $\mathcal{R}^3$. Some authors, for example Jeffreys (1952), and Synge and Schild (1978, p. 128), use the term 'Cartesian tensor' to mean a tensor in this restricted sense.

As will be shown later, the skew-symmetric array $c^{ij} = a^i b^j - a^j b^i$ forms a contravariant tensor of order two quite generally. This array has $n(n-1)/2$ independent non-zero elements. It so happens that if $n = 3$, the three independent elements can be dressed up to masquerade as a vector, at least to the limited extent described above. For many purposes, rotations, angular momentum and similar physical concepts are best thought of as skew-symmetric second-order tensors rather than as vectors. The term 'axial vector' (as opposed to 'polar vector') is sometimes used in physics to distinguish 'vectors' such as $a \times b$ from ordinary vectors in $\mathcal{R}^3$. For example, an electric field is an ordinary (polar) vector in $\mathcal{R}^3$, but a magnetic field is an axial vector (skew-symmetric tensor).

Example 12: Semi-invariants and the affine group. Let $\mathcal{G}$ be the group of $(n+1) \times (n+1)$ matrices of the form

$$g = \begin{pmatrix} 1 & 0 \\ g_0 & g_1 \end{pmatrix},$$

in which $g_0 \in \mathcal{R}^n$, and g_1 is $n \times n$ and non-singular. Let $\mathcal{F}$ be the set of vectors in $\mathcal{R}^{n+1}$ of the form $(1, v)$ with $v \in \mathcal{R}^n$. The action of $\mathcal{G}$ on $\mathcal{F}$

$$\begin{pmatrix} 1 & 0 \\ g_0 & g_1 \end{pmatrix} \begin{pmatrix} 1 \\ v \end{pmatrix} = \begin{pmatrix} 1 \\ g_0 + g_1 v \end{pmatrix} = \begin{pmatrix} 1 \\ \bar{v} \end{pmatrix}.$$

is isomorphic to the action of the general affine group, $GA(\mathcal{R}^n)$ acting on points v in $\mathcal{V} = \mathcal{R}^n$:

$$\bar{v} = g_0 + g_1 v. \tag{1.7}$$

The action of $\mathcal{G}$ on $\mathcal{F}$ is a representation of $GA(\mathcal{R}^n)$ as a sub-group of $GL(\mathcal{R}^{n+1})$, the general linear group acting on $\mathcal{R}^{n+1}$. The set $\mathcal{F}$ is called an affine space in $\mathcal{R}^{n+1}$: it is a translation of its tangent space. The same terminology is used for $\mathcal{V}$, a space subject to affine re-parameterization, even though $\mathcal{V}$ is not obviously a translation of its tangent space in the geometrical sense. Many physical variables such as time, position and velocity, which have no well-defined origin, are subject to affine re-scaling in this sense. The same is true of everyday temperatures recorded in °C or °F.

Now, F is not itself a vector space because it does not include the origin, but it does have a tangent space given by

$$T = \{x_1 - x_2 : x_1, x_2 \in F\} = \{(0, v_1) - (0, v_2) : v_1, v_2 \in \mathcal{V}\}.$$

Evidently, $\mathcal{T}$ is a vector space isomorphic with $\mathcal{R}^n$. In fact, $\mathcal{T}$ may be identified with the set of differences $v_1 - v_2$ with $v_1, v_2 \in \mathcal{V}$. It might seem that $\mathcal{T}$ is the same as $\mathcal{V}$, but it is critically important to distinguish the two for the following reason. The action of the group on a vector $dv \in \mathcal{T}$ is given by

$$d\bar{v} = (\bar{v}_1 - \bar{v}_2) = g_1(v_1 - v_2) = g_1 dv,$$

which, in contrast to (1.7), is linear. If we re-write (1.7) using index notation in the form

$$\bar{v}^i = g^i + g^i_r v^r,$$

we obtain the tensor transformation rule for vectors dv^i in $\mathcal{T}$:

$$d\bar{v}^i = g^i_r dv^r.$$

In other words, although vectors in $\mathcal{V}$ are affected by the change of origin, vectors in the tangent space of $\mathcal{V}$ are unaffected. For that reason, vectors in $\mathcal{T}$ or tensors in $\mathcal{T}^{\otimes k}$ or in $\mathcal{T}^{*\otimes k}$ are sometimes said to be semi-invariant. Cumulants of order two or more are the principal examples of such tensors in statistics.

In the old days when temperature scales were taught in primary schools, the more pedantic teachers, no doubt appreciating the difference between an affine space and its tangent space, insisted on drawing a distinction between degrees centigrade (°C), indicating a temperature, and centigrade degrees (C°), indicating a rise or fall in temperature. While most students eventually came to recognize the distinction, not all of them at the time considered it to be a distinction worth making.

Example 13: Tangent spaces and log likelihood derivatives. Let $\theta = (\theta^1, \ldots, \theta^n)$ be a point in an n-dimensional set or manifold, Θ, and let $l(\theta)$ be a differentiable function defined on Θ. In statistical contexts, θ is a parameter and $l(\theta)$ is a log likelihood function. There is no suggestion that Θ is a vector space, even though the term 'parameter space' is commonly used. By a re-parameterization $\theta \mapsto \phi$ is meant a new coordinate system on Θ, i.e. a set of functions $(\phi^1(\theta), \ldots, \phi^n(\theta))$ whose Jacobian is non-zero. It is usually best to think of the parameter space as a static entity transcending its description by coordinates: a coordinate system then functions like a basis in the sense that it provides a systematic method for identifying points in Θ. The set of such re-parameterizations evidently constitutes a group. The derivative matrix of the transformation is

$$\phi^i_r = \partial\phi^i/\partial\theta^r.$$

The inverse matrix is $\theta^i_r = \partial\theta^i/\partial\phi^r$. Since the transformation is not linear, these matrices are not constant over Θ.

Let $\mathcal{T}_\theta$ be the tangent space at θ, i.e. formally the space spanned by the differential operators $\{\partial/\partial\theta^1, \ldots, \partial/\partial\theta^n\}$ at a specific point θ. To form a mental image of this, it may be helpful to imagine Θ as a two-dimensional surface $x(\theta)$ in $\mathcal{R}^3$ parameterized by (θ^1, θ^2). Then $\partial x/\partial\theta^1, \partial x/\partial\theta^2$ are vectors in $\mathcal{R}^3$, parallel to the tangent plane at θ. The components of such a vector in $\mathcal{T}_\theta$ are denoted by $(d\theta^1, \ldots, d\theta^n)$. Consider by way of example the spherical polar representation of the unit sphere in $\mathcal{R}^3$,

$$x(\theta) = (\cos\theta^1, \sin\theta^1 \cos\theta^2, \sin\theta^1 \sin\theta^2)$$

for $0 < \theta^1 < \pi$, $-\pi < \theta^2 \leq \pi$. The derivative vectors at $\theta^1 = \theta^2 = \pi/4$ are $e_1 = (-1/\sqrt{2}, 1/2, 1/2)$ and $e_2 = (0, -1/2, 1/2)$. These are the basis vectors in $\mathcal{T}_{\pi/4}$, the tangent space at $\theta = (\pi/4, \pi/4)$. A

point in the tangent space is identified by its components relative to this basis, i.e. the components $(d\theta^1, d\theta^2)$ signify the point $e_1 d\theta^1 + e_2 d\theta^2$ in the tangent space. Note that by definition of a space, the tangent space must include the origin, or a point called zero. Thus, $x(\theta)$ is not ordinarily a point in the tangent space. However, in order to enhance the image it is helpful mentally to transport the origin to $x(\theta)$, so that the algebraic definition of tangent space is made to coincide with the geometric notion of tangent plane, which is the set of points $x(\theta) + \lambda^1 e_1 + \lambda^2 e_2$ with $(\lambda^1, \lambda^2) \in \mathcal{R}^2$.

In the dual space of linear functionals on $\mathcal{T}_\theta$, $\alpha = (\alpha_1, \ldots, \alpha_n)$ is identified with the linear functional $\alpha_i d\theta^i$ on $\mathcal{T}_\theta$. Under the action of the group on Θ, basis vectors in $\mathcal{T}_\theta$ are transformed linearly by

$$\frac{\partial}{\partial \theta^i} \mapsto \frac{\partial}{\partial \phi^i} = \frac{\partial}{\partial \theta^r} \theta_i^r .$$

Hence the components transform inversely by the rule

$$d\theta^i \mapsto d\phi^i = \phi_r^i d\theta^r .$$

To draw a parallel with section 1.4.2, $\theta_i^r = K_i^r$ is the matrix of the linear transformation on the space $\mathcal{T}_\theta$, so log-likelihood derivatives transform linearly as covariant tensors as in (1.5). Component vectors $d\theta^i$ are transformed linearly by the inverse matrix $\phi_r^i = L_r^i$, in the manner of contravariant tensors. The only major departure here is that the matrix of the transformation varies from point to point in the parameter space.

0.4 Tensor products

0.4.1 *Examples*

The formation of tensor products is a method for putting together two or more vector spaces to form a new vector space called the tensor product space. The process is much like formal multiplication in the sense that the new space is linear in each of the component spaces. Some familiar examples of tensor product spaces will serve as a guide to motivate the definition.

Example 14: Tensor product of polynomials. Let $\mathcal{V}_1$ be the vector space of affine functions of one variable, x_1. This is a two-dimensional space spanned by the functions $\{1, x_1\}$ on $\mathcal{R}$. Likewise, let $\mathcal{V}_2$ be another vector space of quadratic functions of the variable x_2. Thus $\mathcal{V}_2 = \text{span}\{1, x_2, x_2^2\}$. Then $\mathcal{V}_1 \otimes \mathcal{V}_2$ is the space of functions on the plane spanned by all the formal products

$$\mathcal{V}_1 \otimes \mathcal{V}_2 = \text{span}\{1, x_1, x_2, x_1 x_2, x_2^2, x_1 x_2^2\}.$$

There is a minor technical distinction to be drawn here between the function $x_2 \in \mathcal{V}_2$, a function of one variable, and $x_2 \in \mathcal{V}_1 \otimes \mathcal{V}_2$, a function on the plane which happens to depend only on one of the two components. On balance it seems unnecessary and unhelpful to make such distinctions a part of the notation.

Note that the dimension of the tensor product space is the product of the dimensions of $\mathcal{V}_1$ and $\mathcal{V}_2$.

Example 15: Tensor product of Heaviside functions. Let $H(x)$, the unit Heaviside function, be the indicator function for $x \geq 0$. Define

$$\mathcal{V}_1 = \text{span}\{1, H(x_1)\}, \quad \mathcal{V}_2 = \text{span}\{1, H(x_2)\}.$$

Then $\mathcal{V}_1 \otimes \mathcal{V}_2$, the space of formal products as functions on the plane, is the space of functions that are constant on quadrants. In symbols,

$$\mathcal{V}_1 \otimes \mathcal{V}_2 = \text{span}\{1, H(x_1), H(x_2), H(x_1)H(x_2)\}.$$

Contrast the direct sum space $\mathcal{V}_1 \oplus \mathcal{V}_2$, which is

$$\mathcal{V}_1 \oplus \mathcal{V}_2 = \text{span}\{1, H(x_1), H(x_2)\}.$$

In statistical terminology, if binary factor A is associated with x_1, and factor B with x_2, positive values denoting high levels, then $\mathcal{V}_1 \oplus \mathcal{V}_2$ is the main-effects model $A+B$, whereas $\mathcal{V}_1 \otimes \mathcal{V}_2$ includes the interaction and is written as $A*B$ or $A.B$. By one convention, $A*B$ means $\mathcal{V}_1 \otimes \mathcal{V}_2$ with the basis given above: $A.B$ denotes the same space with quadrant indicator functions as basis vectors.

Example 16: Tensor product of contrast vectors. Let $\mathcal{V}$ be the space of contrast vectors in $\mathcal{R}^3$ spanned by

$$e_1 = (1, 0, -1), \quad \text{and} \quad e_2 = (0, 1, -1).$$

Then the basis vectors of $\mathcal{V} \otimes \mathcal{V} = \mathcal{V}^{\otimes 2}$ are the formal products $e_{ij} = e_i \otimes e_j$. If we regard e_i as a column vector it is simplest to portray $e_{ij} = e_i e_j^T$ as a rank-one matrix.

$$e_{11} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \qquad e_{12} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

$$e_{21} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix}, \qquad e_{22} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Note that e_{11}, e_{22} and $e_{12} + e_{21}$ are symmetric matrices whereas $e_{12} - e_{21}$ is skew-symmetric. The symmetric matrices span a subspace of $\mathcal{V}^2$ called $\text{sym}_2(\mathcal{V})$: the skew-symmetric matrices span a subspace called $\text{alt}_2(\mathcal{V})$. For most statistical purposes, only the symmetric part is relevant.

The same space and subspaces, though with different basis vectors, would be obtained were we to use a different basis for $\mathcal{V}$.

0.4.2 *Definition of tensor product*

Probably the easiest way to envisage a tensor product of two vectors $v_1 \in \mathcal{V}_1$ and $v_2 \in \mathcal{V}_2$ is to think of vectors as functions of one variable. A vector $(\xi_1, \ldots, \xi_n) \in \mathcal{R}^n$ can be expressed as a real-valued function $\xi(\cdot)$ defined on the first n positive integers, whose value at $x = i$ is $\xi(i) = \xi_i$. The tensor product $v_1 \otimes v_2$ is then a function on the plane whose value at (x_1, x_2) is equal to the product $v_1(x_1)v_2(x_2)$.

Definition: Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces. For each $v_1, v_2 \in \mathcal{V}$ and $w_1, w_2 \in \mathcal{W}$, the symbol $v \otimes w$ is defined to be the bi-linear formal product satisfying

$$(\alpha_1 v_1 + \alpha_2 v_2) \otimes w = \alpha_1 v_1 \otimes w + \alpha_2 v_2 \otimes w,$$
$$v \otimes (\alpha_1 w_1 + \alpha_2 w_2) = \alpha_1 v \otimes w_1 + \alpha_2 v \otimes w_2.$$

for all scalars α_1, α_2. Then $\mathcal{V} \otimes \mathcal{W}$ is the vector space of all linear combinations $\sum \alpha_i v_i \otimes w_i$ for $v_i \in \mathcal{V}$ and $w_i \in \mathcal{W}$. If $\{e_1, \ldots, e_m\}$ is a basis in $\mathcal{V}$ and $\{e'_1, \ldots, e'_n\}$ is a basis in $\mathcal{W}$ then the set of mn symbols $\{e_{ij} = e_i \otimes e'_j\}$ is a basis, called a product basis, in $\mathcal{V} \otimes \mathcal{W}$.

0.4.3 *Bi-linear functionals*

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces. A bi-linear functional $y(v, w)$ on $\mathcal{V} \times \mathcal{W}$ is a scalar-valued function, defined for each $v \in \mathcal{V}$ and $w \in \mathcal{W}$, that is linear in v for fixed w, and linear in w for fixed v. In symbols

$$y(\alpha_1 v_1 + \alpha_2 v_2, w) = \alpha_1 y(v_1, w) + \alpha_2 y(v_2, w),$$
$$y(v, \alpha_1 w_1 + \alpha_2 w_2) = \alpha_1 y(v, w_1) + \alpha_2 y(v, w_2)$$

for each $v, v_1, v_2 \in \mathcal{V}$, for $w, w_1, w_2 \in \mathcal{W}$, and for all scalars α_1, α_2. It is clear from the definition that if $y_1(\cdot, \cdot)$ and $y_2(\cdot, \cdot)$ are bi-linear functionals, so also is the linear combination $\alpha_1 y_1 + \alpha_2 y_2$. In other words, the set of bi-linear functionals on $\mathcal{V} \times \mathcal{W}$ is a vector space. It is possible now to give an alternative definition of the tensor product space $\mathcal{V} \otimes \mathcal{W}$ as follows:

Definition: The tensor product of two vector spaces $\mathcal{V}$ and $\mathcal{W}$ is the dual of the space of bi-linear functionals on $\mathcal{V} \times \mathcal{W}$.

If $(v^1, \ldots, v^m)$ are the components of $v \in \mathcal{V}$ and $(w^1, \ldots, w^n)$ are the components of $w \in \mathcal{W}$, then v^i is an example of a linear functional on $\mathcal{V}$ and w^i is an example of a linear functional on $\mathcal{W}$. In fact, these are the dual bases vectors in $\mathcal{V}^*$ and $\mathcal{W}^*$ respectively. Consequently, the product $(v, w) \mapsto v^i w^j$ is an example of a bi-linear functional on $\mathcal{V} \times \mathcal{W}$. In fact these mn bi-linear functionals constitute the dual basis in $\mathcal{V}^* \otimes \mathcal{W}^*$.

0.4.4 *Components of a tensor*

Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be vector spaces with bases $\{e_1, \ldots, e_m\}$ and $\{e'_1, \ldots, e'_n\}$ respectively. Then $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ is a vector space of dimension mn with basis vectors

$$\{e_{ij} = e_i \otimes e'_j; \qquad i = 1, \ldots, m; j = 1, \ldots, n\}.$$

Each point $w \in \mathcal{W}$ has a unique representation as a linear combination of the basis vectors. We write

$$w = w^{ij} e_i \otimes e'_j = w^{ij} e_{ij},$$

making use of the summation convention in which the range of summation for each index is to be understood from the context. Then w^{ij} are the components of the tensor w. It is an obvious abuse of terminology to say that w^{ij}, meaning the array of real numbers, is a (contravariant) tensor, but such abuses are not uncommon.

0.4.5 *Tensor transformation rules*

Let $\{e_1, \ldots, e_m\}$ be a basis in $\mathcal{V}_1$, $\{e'_1, \ldots, e'_n\}$ be a basis in $\mathcal{V}_2$, and let w^{ij} be the components of a tensor or vector in $\mathcal{W} = \mathcal{V}_1 \otimes \mathcal{V}_2$ with respect to the product basis $\{e_{ij} = e_1 \otimes e'_j\}$. In symbols,

$$w = w^{ij} e_{ij}.$$

Consider now the effect of introducing a new basis $\{\bar{e}_i\}$ in $\mathcal{V}_1$ and $\{\bar{e}'\}$ in $\mathcal{V}_2$. We write

$$\bar{e}_r = K^i_r e_i, \qquad \bar{e}'_r = K'^i_r e'_i,$$

defining the new basis vectors in terms of the old. The product basis in $\mathcal{W}$ is thereby transformed from $\{e_{ij}\}$ to $\{\bar{e}_{ij}\}$ given by

$$\bar{e}_{rs} = K^i_r K'^j_s e_i \otimes e'_j = K^i_r K'^j_s e_{ij}. \tag{1.8}$$

Accordingly, if we denote by $\bar{w}^{ij}$ the components of w with respect to the new basis, we have

$$w = \bar{w}^{rs} \bar{e}_{rs} = \bar{w}^{rs} K^i_r K'^j_s e_{ij} = w^{ij} e_{ij}.$$

Thus the components of w with respect to the new basis are

$$\bar{w}^{rs} = L_i^r L_j'^s w^{ij} \tag{1.9}$$

in which L is the matrix inverse of K, and L' is the inverse of K'. This is the tensor transformation rule for the components of a contravariant tensor in $\mathcal{V}_1 \otimes \mathcal{V}_2$.

Components of vectors in the dual space $\mathcal{W}^*$ of linear functionals on $\mathcal{W}$ are denoted by means of subscripts. Their tensor transformation rule is given by (1.8).

In most of the applications that we have in mind only one space $\mathcal{V}$, and its tensor products $\mathcal{V}^{\otimes k}$, is involved. Thus $\mathcal{V}_1 = \mathcal{V}_2$, $m = n$, $K = K'$ and $L = L'$.

The extension of the tensor transformation rules (1.8)and (1.9) for tensors of higher order should be immediately apparent.

0.4.6 *Tensor multiplication and contraction*

Operationally, a tensor is an array of numbers or functionals that obeys a particular multi-linear transformation rule under a change of basis in the relevant space or spaces. Suppose for simplicity that a single space $\mathcal{V}$ is involved, and that the change of basis vectors is given by

$$\bar{e}_i = a_i^r e_r \quad \text{or} \quad e_r = b_r^i \bar{e}_i,$$

where b_r^i is the matrix inverse of a_i^r. These matrices belong to some group relevant to the problem, usually the full linear group, but in appropriate circumstances, the orthogonal group or the symmetric group. An array $\omega = \omega^{ijk}$ whose values in the new coordinate system are given by

$$\bar{\omega}^{ijk} = b_r^i b_s^j b_t^k \omega^{rst}$$

is said to be a contravariant tensor of order three. Technically, ω^{ijk} are the components of a point or vector in $\mathcal{V}^{\otimes 3}$. Likewise, an array γ_{ij} that transforms by the rule

$$\bar{\gamma}_{ij} = a_i^r a_j^s \gamma_{rs}$$

is said to be covariant of order two. The direct product of two tensors is also a tensor. For example,

$$\psi_{rs}^{ijk} = \omega^{ijk} \gamma_{rs}$$

is a tensor of covariant order two and contravariant order three. It obeys the transformation rule for a mixed tensor.

Direct multiplication results in a tensor of increased order. In general, the indices may have different ranges and may refer to quite unrelated spaces. However, when a superscript and a subscript refer to components in a space and its dual, they necessarily have the same range. In such cases it is legitimate to sum over the index, a process known as contraction, resulting in a tensor of lower order. In the preceding example,

$$\phi^i = \psi_{rs}^{irs} = \omega^{irs} \gamma_{rs}$$

is a contravariant vector, or the component vector of a point in $\mathcal{V}$. It is an elementary exercise to show that ϕ^i does obey the transformation rule for component vectors in $\mathcal{V}$.

Tensor calculus is simply the application of these multiplication and contraction rules: it is ordinarily not legitimate, for example, to sum over a single index, or to make two superscripts equal and to sum over the 'diagonal.'

One of the most important cases of tensor contraction occurs when it results in a tensor of order zero, otherwise known as a scalar or an invariant. Such quantities are particularly important because their value is the same in every coordinate system within the group.

0.5 Examples

The following is a brief selection of spaces and tensors that arise in various statistical operations.

Example 17: Moments of random variables. A random variable $X(\omega)$ is conceived of as a function from the outcome space Ω with values in a vector space $\mathcal{V}$. The average $\mu = E(X) = \int X(\omega)\, dP(\omega)$ is a linear combination of vectors in $\mathcal{V}$. Hence $\mu \in \mathcal{V}$. To say the same thing in an equivalent way, the vector with components

$$\mu^r = E(X^r) = \int X^r(\omega)\, dP(\omega)$$

are the components of a vector in $\mathcal{V}$. Likewise, since $X^r(\cdot)$ is a linear functional on $\mathcal{V}$ and $X^r X^s$ is a bi-linear functional, it follows that

$$\mu^{rs} = E(X^r X^s) = \int X^r(\omega) X^s(\omega)\, dP(\omega)$$

are the components of a vector in $\mathcal{V}^{\otimes 2}$.

Usually, we omit mention of the outcome space Ω and write the first three moments of X as follows:

$$\begin{aligned} \kappa^i &= E(X^i), \\ \kappa^{ij} &= E(X^i X^j), \\ \kappa^{ijk} &= E(X^i X^j X^k). \end{aligned}$$

these being the components of vectors in $\mathcal{V}$, $\mathcal{V}^{\otimes 2}$, and $\mathcal{V}^{\otimes 3}$ respectively.

To say the same thing in another way, consider the linear transformation $\bar{X}^i = a^i_r X^r$ as in (1.6). Then the moments of the transformed random variable are

$$\begin{aligned} \bar{\kappa}^i &= E(a^i_r X^r) = a^i_r E(X^r) = a^i_r \kappa^r, \\ \bar{\kappa}^{ij} &= E(a^i_r a^j_s X^r X^s) = a^i_r a^j_s E(X^r X^s) = a^i_r a^j_s \kappa^{rs}, \end{aligned}$$

and so on. These are the transformation rules for the components of contravariant tensors in $\mathcal{V}$, $\mathcal{V}^{\otimes 2}$ and so on.

Example 18: Covariance matrix and the affine group. Let $X = (X^1, \ldots, X^n)$ be the components of a random vector in an affine space $\mathcal{V} = \mathcal{R}^n$. We consider $\mathcal{V}$ to be an affine space in the sense that it has no definite origin, or in the sense that it is subject to affine re-parameterization. One way to define the covariance matrix of X is as as follows. Let X_1 and X_2 be independent copies of X, and let $dX = X_1 - X_2$. Then the components of the covariance matrix of X are given by

$$\kappa^{i,j} = \tfrac{1}{2} E(dX^i\, dX^j).$$

In the terminology of Example 4 in section 1.4.3, dX^i are the components of a point in $\mathcal{T}$, the tangent space of $\mathcal{V}$. Since the expectation operator is linear, and $dX^i\, dX^j$ is bi-linear, it follows that $\kappa^{i,j}$ are the coordinates of a point (contravariant tensor) in $\mathcal{T}^{\otimes 2}$. By contrast, the moment matrix with components $\kappa^{ij} = E(X^i X^j)$, is not a tensor in $\mathcal{T}^{\otimes 2}$. The previous example shows that κ^{ij} are the components of a tensor in $\mathcal{V}^{\otimes 2}$, but only if $\mathcal{V}$ is regarded as a vector space rather than an affine space.

We now check the preceding argument in the standard way. The covariance matrix of X has components $\kappa^{i,j}$ defined by

$$\kappa^{i,j} = E(X^i X^j) - E(X^i)E(X^j) = \kappa^{ij} - \kappa^i \kappa^j.$$

Consider now the general affine group

$$\bar{X}^i = a^i + a^i_r X^r$$

acting on $\mathcal{V}$. In order to show that $\kappa^{i,j}$ are the components of a vector in $\mathcal{T}^{\otimes 2}$ we need to show that the covariance matrix of the transformed variable satisfies $\bar{\kappa}^{i,j} = a^i_r a^j_s \kappa^{r,s}$. From the linearity of the expectation operator, we have

$$\begin{aligned}\bar{\kappa}^i &= E(\bar{X}^i) = a^i + a^i_r \kappa^r, \\ \bar{\kappa}^{ij} &= E(\bar{X}^i \bar{X}^j) = a^i a^j + a^i a^j_r \kappa^r + a^j a^i_r \kappa^r + a^i_r a^j_s \kappa^{rs}.\end{aligned}$$

Then the product of the mean vectors is given by

$$\bar{\kappa}^i \bar{\kappa}^j = a^i a^j + a^i a^j_r \kappa^r + a^i a^j_r \kappa^r + a^i_r a^j_s \kappa^r \kappa^s.$$

Finally, the difference gives

$$\bar{\kappa}^{i,j} = \bar{\kappa}^{ij} - \bar{\kappa}^i \bar{\kappa}^j = a^i_r a^j_s (\kappa^{rs} - \kappa^r \kappa^s) = a^i_r a^j_s \kappa^{r,s},$$

which is the required transformation rule for tensors in $\mathcal{T}^{\otimes 2}$.

Example 19: Metric tensor in Euclidean space. Let $x = (x^1, \ldots, x^n)$ be the coordinates of a point in an n-dimensional affine space $\mathcal{E}$. To say that the space is Euclidean is to say that the squared distance between points x and y in $\mathcal{E}$ is given by the quadratic form

$$ds^2 = d^i d^j g_{ij}, \tag{1.10}$$

where $d^i = x^i - y^i$. The coefficients g_{ij} are the components of a symmetric positive definite matrix, called the metric or metric tensor in $\mathcal{E}$.

Now, d^i are the components of a vector in $\mathcal{T}$, the tangent space of $\mathcal{E}$. In order for ds^2 to be invariant under affine transformation it is necessary for g_{ij} to be the components of a tensor in $\mathcal{T}^* \otimes \mathcal{T}^*$. In other words, g_{ij} are the components of a covariant tensor of order two.

The most natural example of a metric tensor is the inner product matrix $G_{ij} = \langle e_i, e_j \rangle$, where $\{e_i\}$ are the basis vectors in $\mathcal{T}$. Since $\langle \cdot, \cdot \rangle$ is a symmetric bi-linear functional on $\mathcal{T}$, it follows automatically that G_{ij} are the components of a symmetric covariant tensor, called the fundamental metric tensor in $\mathcal{T}$, or in $\mathcal{E}$.

Example 20: Metric tensor in Riemannian space. In a Riemannian space, the metric tensor g_{ij} is not constant over the space, but varies continuously from point to point in the manifold E. Then (1.10) gives the infinitesimal squared distance between neighbouring points x and $x + dx$ in terms of the metric g_{ij} at x. Under arbitrary differentiable re-parameterizations of the manifold, dx^i are the components of a vector in the tangent space, $\mathcal{T}_x$, of E at x. By the argument given above for Euclidean spaces, since ds^2 is invariant, g_{ij} must be the components of a tensor in $\mathcal{T}_x^* \otimes \mathcal{T}_x^*$.

Example 21: Fisher information matrix. Let $l(\theta; y)$ denote the log likelihood function defined on $\Theta \times \mathcal{Y}$, where Θ denotes the parameter space, and $\mathcal{Y}$ denotes the sample space. The components of the observed and expected Fisher information matrices are as follows:

$$I_{ij} = -\frac{\partial^2 l(\hat{\theta}; y)}{\partial \theta^i \partial \theta^j}; \qquad \mathcal{I}_{ij} = -E_\theta\Big(\frac{\partial^2 l(\theta; Y)}{\partial \theta^i \partial \theta^j}\Big).$$

Under differentiable re-parameterizations $\theta \mapsto \phi$, component vectors in the tangent space of Θ undergo linear transformation via

$$d\theta^i \mapsto d\phi^i = \phi^i_r d\theta^r,$$

where $\phi^i_r = \partial \phi^i / \partial \theta^r$ is the Jacobian matrix at θ. It is a straightforward exercise to check that the components of I and $\mathcal{I}$ undergo linear transformation to

$$\bar{I}_{ij} = \hat{\theta}^r_i \hat{\theta}^s_j I_{rs} \quad \text{and} \quad \bar{\mathcal{I}}_{ij} = \theta^r_i \theta^s_j \mathcal{I}_{rs}, \tag{1.11}$$

where $\theta_i^r = \partial\theta^r/\partial\phi^i$, and $\hat{\theta}_i^r$ is the same matrix evaluated at $\hat{\theta}$. To show that I is a tensor, it is necessary to assume that $\partial l(\theta; y)/\partial\theta^i = 0$ at $\hat{\theta}$. To show that $\mathcal{I}$ is a tensor, it is necessary to assume that $E_\theta(\partial l(\theta; Y)/\partial\theta^i) = 0$. As a consequence of (1.11), we say that $\mathcal{I}_{ij}$ are the components of a covariant tensor in $\mathcal{T}_\theta^* \otimes \mathcal{T}_\theta^*$, the tensor product of the dual of the tangent space at θ: I_{ij} are the components of a covariant tensor in $\mathcal{T}_{\hat{\theta}}^* \otimes \mathcal{T}_{\hat{\theta}}^*$, the tensor product of the dual of the tangent space at $\hat{\theta}$. So, although both information matrices are symmetric covariant tensors, they are not covariant tensors in the same tangent space.

Note that $\mathcal{I}$ is a function on Θ, in fact a covariant tensor field on Θ. As such, Θ can be regarded as a Riemannian space with $\mathcal{I}$ as metric tensor.

Example 22: Kronecker delta. Kronecker's delta is the symbol δ_i^j, the identity transformation from a vector space to itself. Consider, however, the symbols δ^{ij}, δ_{ij} taking the value 1 if $i = j$ and zero otherwise. What does it mean for these to be the components of a tensor, presumably in the tensor product spaces $\mathcal{V} \otimes \mathcal{V}$ and $\mathcal{V}^* \otimes \mathcal{V}^*$? Unless these are the components relative to a very particular coordinate system, for example the orthogonal coordinate system in $\mathcal{R}^3$ with equal units of distance in each coordinate direction, it is to be understood that δ^{ij} are the components in all coordinate systems. Under linear transformation in $\mathcal{V}$ the components must change to $\bar{\delta}^{ij}$, also taking the value 1 if $i = j$ and zero otherwise. In symbols, we must have

$$\bar{\delta}^{ij} = g_r^i g_s^j \delta^{rs} = \delta^{ij},$$

for all matrices g in the group $\mathcal{G}$. This identity holds only for the orthogonal group, and in that restricted sense only, δ^{ij} are the components of a tensor in $\mathcal{V}^{\otimes 2}$. In the same sense δ_{ij} are the components of a tensor in the dual space. Such tensors, whose components are the same under all relevant coordinate transformations, are said to be isotropic (Jeffreys, 19??).

In real space $g_{ij} = \langle e_i, e_j \rangle$ is the inner product or metric tensor on $\mathcal{V}$. There does exist a basis for which $g_{ij} = \delta_{ij}$, namely any standard orthogonal basis with equal units in each of the coordinate directions. From the viewpoint of tensor calculus, it is usually best not to place any restrictions on the basis vectors. In other words, if the calculations involve a metric space, it is best to use a symbol g_{ij}, recognizing that these are the components of a tensor in $\mathcal{T}^* \otimes \mathcal{T}^*$, the tensor product of the dual of the tangent space of $\mathcal{V}$ with itself. Explicit numerical computations may subsequently exploit the orthogonality of the basis vectors, but it is usually a tactical mistake to impose such restrictions at an early stage in the algebra.

Example 23: Extended Kronecker delta. Let δ^{ijk}, δ^{ijkl} take the value 1 if all indices are equal, and zero otherwise. These are isotropic tensors, but only under the symmetric group (permutation of coordinates or indices), which is a discrete sub-group of the orthogonal group on $\mathcal{R}^n$.

There are occasions when this group is of statistical interest, for example when we deal with symmetric functions of independent and identically distributed random variables. From an operational or notational point of view, these extended Kronecker symbols should not appear unless the symmetric group is the relevant group of symmetries on $\mathcal{V}$.

0.6 Inner products and conjugate functionals

Notions of magnitude, direction and orthogonality, familiar from our everyday experience in 'real space,' are absent from the standard definition of a vector space. To include this additional structure, we define an inner product. An inner product on a real vector space $\mathcal{V}$ is a real-valued symmetric bi-linear functional satisfying the following conditions.

$$\begin{aligned} \langle x, y \rangle &= \langle y, x \rangle; \\ \langle \alpha_1 x_1 + \alpha_2 x_2, y \rangle &= \alpha_1 \langle x_1, y \rangle + \alpha_2 \langle x_2, y \rangle; \\ \langle x, x \rangle &\geq 0; \\ \text{if} \quad \langle x, x \rangle = 0, &\quad \text{then} \quad x = 0. \end{aligned}$$

As usual, x_1, x_2 and y are elements of $\mathcal{V}$, and α_1, α_2 are real scalars.

These definitions are the same as those used by Halmos (1958). Stone (1987) eschews the final positive definiteness condition.

In terms of the component vectors, an inner product has the form

$$\langle x,\, y\rangle = x^i y^j g_{ij},$$

where $g_{ij} = \langle e_i,\, e_j\rangle$, the fundamental metric tensor on the space, is positive definite.

For any linear functional A on $\mathcal{V}$, there exists an associated linear functional A^* such that, for all x, y,

$$\langle Ax,\, y\rangle = \langle x,\, A^* y\rangle.$$

We say that A^* is the *conjugate linear functional* of A. By symmetry,

$$\langle x,\, A^* y\rangle = \langle A^* y,\, x\rangle = \langle y,\, A^{**} x\rangle = \langle A^{**} x,\, y\rangle,$$

so that $A^{**} = A$. In matrix notation, if $\langle x,\, y\rangle = x^T G y$, then

$$\langle Ax,\, y\rangle = (Ax)^T G y = x^T A^T G y = x^T G (G^{-1} A^T G y) = \langle x,\, A^* y\rangle$$

so the conjugate matrix of A is $A^* = G^{-1} A^T G$. In particular, if G is proportional to the identity, the conjugate matrix is the transposed matrix.

Using index notation, and denoting the components of the linear transformation Ax by $a^r_i x^i$, the components of the conjugate transformation are

$$a^{*r}_{\ i} = g^{rs} g_{ij} a^j_s$$

Unlike the adjoint matrix of section 1.4.4, the conjugate requires a non-trivial transformation, a tensor-preserving transposition of subscripts and superscripts.

A matrix that is self-conjugate satisfies the condition $A^T G = GA$, so that GA is symmetric.

0.7 Generalized inverse matrices

A square matrix A of full rank has a unique inverse usually denoted by A^{-1}. If A does not have full rank, any matrix A^- satisfying

$$AA^-A = A \tag{1.12}$$

is said to be a generalized inverse of A. Apart from the full-rank case, there are infinitely many matrices A^- satisfying the defining property of a generalized inverse. However, certain derived properties or derived vectors in statistical applications are independent of the choice of generalized inverse. In particular, since $(AA^-)(Ax) = Ax$, it follows that AA^- is the identity on the range of A. Consequently, the rank of AA^- is at least equal to the dimension of the range of A, i.e. $\mathrm{rank}(AA^-) \geq \mathrm{rank}(A)$. But the rank of a matrix product cannot exceed the rank of either matrix. It follows that $\mathrm{rank}(AA^-) = \mathrm{rank}(A) \leq \mathrm{rank}(A^-)$. Moreover, from the defining property, we have $(AA^-)(AA^-) = AA^-$. In other words AA^- is idempotent, a projection matrix whose eigenvalues are either zero or one. The same is true of A^-A. Consequently,

$$\mathrm{tr}(AA^-) = \mathrm{rank}(AA^-) = \mathrm{rank}(A) \leq \mathrm{rank}(A^-),$$

independently of the choice of generalized inverse.

If B is a full-rank matrix, then $B^{T-1} A^- B^{-1}$ is a generalized inverse of BAB^T because

$$BAB^T (B^{T-1} A^- B^{-1}) BAB^T = BAA^-AB = BAB^T,$$

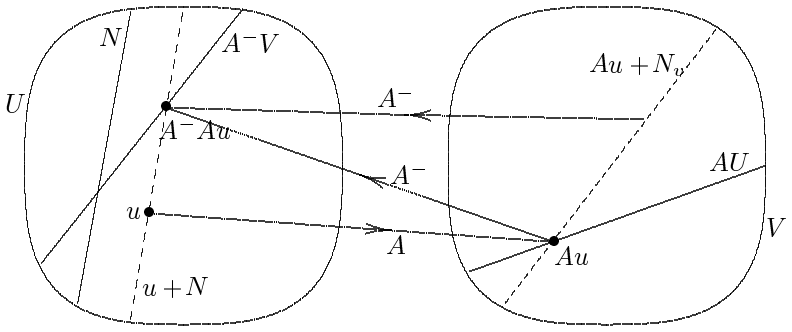

Figure 1.2. *Stone diagram illustrating the definition of an annihilating generalized inverse of a linear transformation from U to V. The range of A is the solid line marked AU in V: the range of A^- is the solid line marked A^-V in U. The solid line N is the null space of A in U. The dashed line $u + N$, the translate of the null space by u, is the pre-image of Au. In the inverse transformation, all points in $Au + N_v$ (dashed line in V), the translate of the null space of A^-, are transformed by A^- to the same point in the pre-image. In the Moore-Penrose inverse, A^-V is orthogonal to N in U, and AU is orthogonal to N_v in V.*

and the defining property is satisfied.

Reverting now to index notation, and supposing A to be symmetric, we write a^{ij} for the components of A and a_{ij} for the components of A^-. The rank of A is then given by $a^{ij}a_{ij}$, the trace of the matrix product. Furthermore, if a^{ij} are the components of a contravariant tensor, the equation above shows that a_{ij} are the components of a covariant tensor. Consequently, $a^{ij}a_{ij}$ is the contraction of a tensor product, or a tensor of order zero, and is invariant.

These results are entirely independent of the choice of generalized inverse. Occasionally, however, it is convenient or even necessary to choose a generalized inverse with the egalitarian property that A is a generalized inverse of A^-. In symbols,

$$A^-AA^- = A^-. \tag{1.13}$$

This property ensures that the two matrices have equal rank. A matrix satisfying the first property is called a *minimal inverse*: a matrix satisfying both properties is called an *annihilating inverse* (Stone, 1987), or a *reflexive inverse* (Rao, 1973).

It may be helpful to explain in coordinate-free geometrical language the rationale for conditions (1.12) and (1.13). Let A be a linear transformation between two vector spaces $\mathcal{U}$ and $\mathcal{V}$ (Fig. 1.1). The null space, $\mathcal{N}$, is the set of vectors in $\mathcal{U}$ such that $Au = 0$. Consequently, all vectors in the set $u + \mathcal{N}$, and only those vectors, are mapped to the point $y = Au$ in the range. The set of points that are mapped to y by A is called the *pre-image of y in $\mathcal{U}$*. Any matrix A^- that transforms y to a point in the pre-image of y is a generalized inverse of A In symbols, the minimal requirement for a generalized inverse is that, for each u, $A^-y = A^-Au$ should lie in $N + u$. Consequently $AA^-Au = Au$, and (1.12) follows.

The minimal property is a condition on the action of A^- on the range of A in $\mathcal{V}$. It says nothing about the action of A^- on any v that is not in the range. The annihilating condition requires A^- to have a null space $N_v \subset \mathcal{V}$ that is complementary to the range. In other words, each $v \in \mathcal{V}$ has a unique decomposition $v = v_0 + y$ with $A^-v_0 = 0$, and $y = Au$ in the range of A. Consequently, the set of points $y + N_v$ is transformed by A^- to a single point A^-y in the pre-image of y. In symbols, for any v_0 in the null space, we have $A^-AA^-v_0 = A^-v_0 = 0$, while for any $y = Au$ in the range, the minimal property gives

$$A^-AA^-y = A^-AA^-Au = A^-Au = A^-y.$$

Consequently, $A^-AA^-v = A^-v$ for all $v \in \mathcal{V}$, and (1.13) follows.

Figure 1.1 illustrates these relationships by means of a Stone diagram in which vector spaces are represented by cartouches, subspaces by solid lines, translated subspaces by dashed lines, and transformations by directed lines between two cartouches. Stone diagrams have a certain runic quality that makes them appear impenetrable at first glance. Unlike Venn diagrams, they require careful study to understand, and considerable practice to construct. But, once mastered, Stone diagrams provide the kind of information that is not readily extracted from matrix formulae.

For statistical work, minimal and annihilating inverses are ordinarily the only types of generalized inverse required. The geometrical description makes it clear that it is always possible to choose a generalized inverse satisfying both (1.12) and (1.13), but these conditions do not define a unique inverse. It is of some interest to investigate what further conditions are required to ensure uniqueness. Provided that $\mathcal{U}$ and $\mathcal{V}$ are inner product spaces, two further conditions are enough. Let the inner products be given by $\langle u_1, u_2\rangle = u_1^T \mathcal{W}_u u_2$ and $\langle v_1, v_2\rangle = v_1^T \mathcal{W}_v v_2$, where W_u and W_v are real positive definite symmetric matrices. The Moore-Penrose inverse requires that the null space of A be orthogonal to the range of A^-, and, symmetrically, that the null space of A^- be orthogonal to the range of A. For any u in $\mathcal{U}$, A^-Au lies in the range of A^-, and $u - A^-Au$ lies in the null space of A. The first Moore-Penrose orthogonality condition is thus $\langle A^-Au, u - A^-Au\rangle = 0$ for all $u \in \mathcal{U}$. In matrix notation, we have

$$u^T(A^-A)^T W_u(I - A^-A)u = 0.$$

Consequently, the matrix of this quadratic form must be skew-symmetric.

$$(A^-A)^T W_u - (A^-A)^T W_u A^-A = -W_u A^-A + (A^-A)^T W_u A^-A.$$

Pre-multiplication by $(A^-A)^T$ and post-multiplication by A^-A give

$$(A^-A)^T W_u = (A^-A)^T W_u A^-A = W_u A^-A.$$

In other words, A^-A is required to be a self-conjugate projection matrix on $\mathcal{U}$. The same argument with A and A^- reversed gives the second orthogonality condition in $\mathcal{V}$:

$$(AA^-)^T W_v = (AA^-)^T W_v AA^- = W_v AA^-.$$

In particular, the familiar least-squares projection matrices $P_X = X(X^TWX)^{-1}X^TW$ and $I - P_X$, are self-conjugate and self-inverse in the Moore-Penrose sense. These matrices are considered as transformations from the sample space $\mathcal{U}$ to itself, the inner product matrix being W.

Different choices of inner product matrices on the two spaces yield different Moore-Penrose inverses, a point that is rarely emphasized. In common with most authorities, Rao (1973, p. 26) and Kruskal (1975) take W_u and W_v to be the identity matrices. Stone's definition coincides with ours.

Example 24: Multinomial covariance matrix. Let $\mathcal{V}$ be an inner product space with metric tensor $W = \{w_{ij}\}$, symmetric and positive definite, and let $\mathcal{U} = \mathcal{V}^*$ be the dual of $\mathcal{V}$. The dual inner product matrix in $\mathcal{V}^*$ is W^{-1}. In statistical applications, $W^{-1} = \Sigma$ is usually a covariance matrix. A contravariant tensor with components a^{ij} may be regarded as the components of a linear transform from $\mathcal{U} = \mathcal{V}^*$ to $\mathcal{V}$. Thus $W_u = \Sigma$ and $W_v = \Sigma^{-1}$. Suppose that A is given in matrix notation by

$$A = W_u(I - X(X^TW_uX)^{-1}X^TW_u) = W_u(I - P),$$

where X is a given matrix whose columns are linearly independent vectors in $\mathcal{U}$. Then the Moore-Penrose inverse is

$$A^- = (I - P)W_v$$

In particular, if $W_u = \mathrm{diag}\{\pi_1, \ldots, \pi_n\}$ with $\sum \pi_j = 1$, and $X = 1$, the constant vector, then A is the multinomial covariance matrix whose Moore-Penrose inverse is

$$\mathrm{diag}\{1/\pi_1, \ldots, 1/\pi_n\} - 11^T.$$

In this instance, the null space of A is the span of 1, while the null space of A^- is the span of the vector π.

0.8 Orthogonal projection

Let $\mathcal{V}$ be a real vector space with inner product $\langle \cdot, \cdot \rangle$, symmetric and linear in each argument. By the orthogonal projection of $y \in \mathcal{V}$ on to the subspace $\mathcal{V}_1 \subset \mathcal{V}$, we mean a vector $\hat{y} \in \mathcal{V}_1$ satisfying $\langle y - \hat{y}, \hat{y} \rangle = 0$. In order to construct this projection, or the components of the projection with respect to a basis $\{e_r\}$ in $\mathcal{V}_1$, we proceed as follows.

Denote by $g_{rs} = \langle e_r, e_s \rangle$ the fundamental metric tensor on $\mathcal{V}_1$. As usual, the inverse matrix is denoted by g^{rs}. Define $\hat{y}_r = \langle e_r, v \rangle$. These are the components of a vector in $\mathcal{V}_1^*$. Further, $\hat{y}^r = g^{rs}\hat{y}_s$ are the components of $\hat{y} \in \mathcal{V}_1$ with respect to the basis $\{e_r\}$. In symbols,

$$\hat{y} = \hat{y}^r e_r = g^{rs} e_r \langle e_s, y \rangle$$

is the projection of y on to $\mathcal{V}_1$.

To write the same thing using matrix notation, suppose that $\langle x, y \rangle = x^T W y$, where W is $n \times n$ and positive definite. Suppose, in addition, that e_r is identified with the rth column of the $n \times p$ matrix X, so that $g_{rs} = (X^T W X)_{rs}$. Then $\hat{y}_r = (X^T W y)_r$, $\hat{y}^r = ((X^T W X)^{-1} X^T W Y)_r$, and $\hat{y} = X(X^T W X)^{-1} X^T W Y$, a familiar expression from linear models.

Note that $X^- = (X^T W X)^{-1} X^T W$ satisfies the minimal condition for a generalized inverse of X, so we may write $\hat{y} = XX^- y$ with XX^- as an explicit projection matrix. This particular choice of generalized inverse yields the orthogonal, or self-conjugate, projection: other choices of generalized inverse yield non-orthogonal projections on to the same space.

The preceding analysis assumes that the columns of X are linearly independent so that $X^T W X$ is non-singular. For a variety of reasons, this condition is frequently not satisfied in statistical work. In order to make progress, it is natural to construct the matrix

$$P = X(X^T W X)^- X^T W$$

for some choice of generalized inverse, with the intention that PY should be the required projection on to the column space of X. It is straightforward to check that P is a projection matrix ($P^2 = P$), if and only if the generalized inverse is annihilating. Furthermore, if $X^T W X$ has the same rank as X, all annihilating inverses yield the same projection matrix, and hence the same fitted value or least-squares projection.

Example 25: Estimating functions. Let Y be a random vector in an n-dimensional affine space $\mathcal{V}$. Define $\mu = E(Y)$, $\Sigma = \mathrm{cov}(Y)$, and $W = \Sigma^{-1}$, where Σ is taken to be positive definite. A model specifies that μ lies in a p-dimensional manifold $\mu(\theta) \subset \mathcal{V}$, with θ varying in a p-dimensional parameter set Θ. In general, Σ depends on μ, and hence on θ. Now, $Y - \mu$ is a vector in $\mathcal{T}$, the tangent space of $\mathcal{V}$, and W is a metric tensor on $\mathcal{T}$. Since W depends on μ, we denote by $\langle \cdot, \cdot \rangle_\theta$, the inner product using the metric tensor W evaluated at $\mu(\theta)$.

Let $e_r = \partial\mu/\partial\theta^r$ for $r = 1, \ldots, p$. In matrix notation, $D = (e_1, \ldots, e_p)$ is the derivative matrix of the mean-value vector with respect to the parameters at θ. These are the natural basis vectors in $\mathcal{T}_\theta$, the tangent space at θ, which may be regarded as a subspace of $\mathcal{T}$. The vector $U \in \mathcal{T}_\theta^*$ with components

$$U_r = \langle e_r, Y - \mu \rangle_\theta = (D^T W (Y - \mu))_r$$

is known as the quasi-likelihood estimating function, or quasi-likelihood score vector. The root of the equation $U_r(\hat{\theta}) = 0$ is the quasi-likelihood estimator: if it exists, it may not be unique. Note that $g^{rs}U_s$ are the components of the projection of $Y - \mu$ on to $\mathcal{T}_\theta$ with respect to the basis $\{e_r\}$. This is the first Newton step towards $\hat{\theta}$.

0.9 Mixed tensors

0.9.1 *Linear transformation of vector spaces*

The space of linear transformations from $\mathcal{U}$ of dimension n to $\mathcal{V}$ of dimension m is the set of functions g on $\mathcal{U}$ with range in $\mathcal{V}$, satisfying

$$g(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 g(x_1) + \alpha_2 g(x_2).$$

Evidently, if g_1 and g_2 are in this set, any linear combination $\alpha_1 g_1 + \alpha_2 g_2$ is also in the set. Consequently the set of such transformations constitutes a vector space, in fact a vector space of dimension mn. The components of g with respect to the basis vectors in $\mathcal{U}$ and $\mathcal{V}$ form a $m \times n$ matrix $\{g^i_r\}$: the components of $y = g(x)$ are given by

$$y^i = g^i_r x^r.$$

Thus g^i_r are the components of the mixed tensor g in the space $\mathcal{U} \otimes \mathcal{V}^*$.

0.9.2 *Factorial design and Yates's algorithm*

Consider a laboratory experiment in which yield is observed at each of four temperatures, (factor A), each of three concentrations of solute, (factor B), and for each of two catalysts, (factor C). The incidence bases for factor A, B and C are

$$\begin{aligned} A &= \text{span}\{e_1 = \delta_{i1},\, e_2 = \delta_{i2},\, e_3 = \delta_{i3},\, e_4 = \delta_{i4}\}, \\ B &= \text{span}\{e'_1 = \delta_{j1},\, e'_2 = \delta_{j2},\, e'_3 = \delta_{j3}\}, \\ C &= \text{span}\{e''_1 = \delta_{k1},\, e''_2 = \delta_{k2}\}. \end{aligned}$$

To say that Y^{ijk} is the observation with temperature at level i, concentration at level j, and catalyst at level j, is to say that Y^{ijk} are the components of a vector Y in $A \otimes B \otimes C$ relative to the product basis $e_{ijk} = e_i \otimes e'_j \otimes e''_k$. In symbols, $Y = Y^{ijk} e_{ijk}$.

In comparative experiments, which are concerned with contrasts rather than levels, it is frequently required to transform to a new basis of orthogonal contrasts for A, such as the polynomial contrasts

$$\begin{pmatrix} \bar{e}_1 \\ \bar{e}_2 \\ \bar{e}_3 \\ \bar{e}_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}.$$

We write this using index notation as $\bar{e}_r = K^j_r e_j$, with inverse transform $e_j = L^r_j \bar{e}_r$. Note that $KK^T = \Lambda = \text{diag}\{4, 20, 4, 20\}$, so that $L = K^T \Lambda^{-1}$, or

$$L^T = \begin{pmatrix} 1/4 & & & \\ & 1/20 & & \\ & & 1/4 & \\ & & & 1/20 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{pmatrix}.$$

The usual polynomial contrast matrix for factor B, giving $\bar{e}'$ in terms of e' is

$$K' = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix}.$$

Since the rows are orthogonal, the inverse matrix, $L' = K'^T \Lambda'^{-1}$, can be obtained in the same way as for factor A.

The components of Y with respect to the contrast basis $\bar{e}_{rst} = \bar{e}_r \otimes \bar{e}'_s \otimes \bar{e}''_t$ are $Y = \bar{Y}^{rst}\bar{e}_{rst}$, where

$$\bar{Y}^{rst} = L^r_i L'^s_j L''^t_k Y^{ijk}. \tag{1.14}$$

These are the standardized factorial contrasts and interactions that emerge at the third and final step in the standard implementation of Yates's algorithm. The unstandardized or raw contrasts are those obtained by omitting the 'divisors' Λ, Λ' and Λ'' from the definitions of L, L' and L'', i.e. by replacing L^T by K rather than $\Lambda^{-1}K$. Each matrix multiplication in (1.14) represents one step of Yates's algorithm.

In the present notation, i, j, k refer to factor levels, whereas r, s, t refer to factor contrasts. Thus $\bar{Y}^{111}$ is the mean response, often denoted by I, $\bar{Y}^{121}$ is the linear effect of solute, usually denoted by B_L, and $\bar{Y}^{312}$ is the quadratic by linear interaction of factors A and C, usually denoted by $A_Q C_L$.

The number of arithmetic operations required to compute the contrast vector in (1.14) is $4n+3n+2n$ multiplications and a similar number of additions. Here $n = 4\times3\times2$ is the total number of observations or factor combinations. In general, if there are p factors with $n_1, \ldots, n_p$ levels, the number of operations required by Yates's algorithm is $n(n_1+\cdots+n_p)$. This number is considerably fewer than n^2, the number of operations required to multiply a vector by an unstructured $n \times n$ matrix. In particular, if $n = 2^p$, the number of operations required is $2n\log_2(n)$.

For further details, see Good (1958) or Takemura (198?).

0.9.3 *Fast Fourier transform*

The discrete Fourier transform F_j of a sequence of numbers x_t is a linear transformation given by

$$nF_j = \sum_{t=0}^{n-1} x_t W^{jt},$$

where $W = \exp(2\pi i/n)$ is a complex nth root of unity, and the superscript on W is a power. The fast Fourier transform is a computational device modelled on Yates's algorithm. In cases where n is a composite number, $n = n_1 n_2 \cdots n_p$, the time index is represented as a multi-index, so x_t becomes $x_{t_1 t_2 \ldots t_p}$. A similar indexing system is used for the sequence of Fourier coefficients. Although the multi-index representation of the discrete Fourier transform is not quite so simple as (1.14), so-called twiddle factors can be introduced to achieve similar gains in computational efficiency. For details of the algorithm, see Bloomfield (1976, ch. 4).

0.10 Cosets, quotient spaces and annihilators

0.10.1 *Quotient space*

Let $\mathcal{V}$ be a vector space, and let $\mathcal{X}$ be a subspace of $\mathcal{V}$. A subspace $\mathcal{Y}$ is said to be *complementary* to $\mathcal{X}$ in $\mathcal{V}$ if $\mathcal{X}$ and $\mathcal{Y}$ are disjoint, $(\mathcal{X} \cap \mathcal{Y} = \{0\})$, and $\mathcal{Y} + \mathcal{X} = \mathcal{V}$. For a given subspace $\mathcal{X}$ there are, usually, infinitely many complementary subspaces. Without some additional structure, there is no natural way to select one of these complements for preferential treatment. There is, however, a natural way of patching together all these complementary subspaces to form a new vector space that, for all practical purposes, gives a unique complement of $\mathcal{X}$ in $\mathcal{V}$.

For any $v \in \mathcal{V}$, $v + \mathcal{X}$ denotes the set of vectors $v + x$ with $x \in \mathcal{X}$, called a *coset* of $\mathcal{X}$, or the translate of $\mathcal{X}$ by v. Unless $v \in \mathcal{X}$, this is not a subspace because it does not include zero. Note that if $v_1 + \mathcal{X} = v_2 + \mathcal{X}$, it does not follow that v_1 and v_2 are equal. It does, however, follow that $v_1 - v_2 \in \mathcal{X}$, so we may say that '$v_1 = v_2$ modulo $\mathcal{X}$.' The cosets thus partition $\mathcal{V}$ into non-overlapping subsets or equivalence classes.

Consider now the set whose elements are the cosets of $\mathcal{X}$ in $\mathcal{V}$. Addition of two cosets is defined by addition of the elements, and scalar multiplicaton is defined likewise. Thus

$$(v_1 + \mathcal{X}) + (v_2 + \mathcal{X}) = v_1 + v_2 + \mathcal{X} + \mathcal{X} = v_2 + v_2 + \mathcal{X},$$
$$\alpha(v_1 + \mathcal{X}) = \alpha v_1 + \alpha\mathcal{X} = \alpha v_1 + \mathcal{X}$$

for any scalar α. The cosets thus form a vector space, $\mathcal{X}$ itself being the zero element. This vector space is called the *quotient space*, and is denoted by the symbol $\mathcal{V}/\mathcal{X}$. The dimension of the quotient space is $\dim(\mathcal{V}) - \dim(\mathcal{X})$.

Example 26: Residual vector in linear models. Consider the linear model, written in conventional matrix notation as

$$Y = X\beta + \epsilon,$$

where X is a given matrix whose columns span a subspace $\mathcal{X}$ of $\mathcal{V} = \mathcal{R}^n$, and ϵ is a random variable in $\mathcal{R}^n$, whose distribution depends on unknown parameters, θ. Let P be any projection matrix having range $\mathcal{X}$, and let $Q = I - P$ be the complementary projection having null space $\mathcal{X}$. For example, we could take $P = X(X^TX)^{-1}X^T$ or $P' = X(X^TWX)X^TW$ for any positive definite matrix W. Then $R = QY$ is a residual vector having a distribution independent of β, but dependent on θ. All residuals are equivalent in the sense that they contain the same information. For example, the residual $R' = Q'Y$ can be obtained from $R = QY$ by the linear transformation $R' = Q'R$, and conversely $R = QR'$. Note that, although R' and R are vectors in different spaces, the difference $R' - R$ is necessarily a linear combination of the columns of X, i.e. an element of $\mathcal{X}$. For this reason, it is generally preferable to regard the residual vector as an element of the quotient space $\mathcal{V}/\mathcal{X}$ rather than an element of any particular subspace complementary to $\mathcal{X}$ in $\mathcal{V}$.

In fact, it is technically not necessary to construct a projection matrix or form a residual vector at all: we may talk simply of the distribution of the random variable $Y + \mathcal{X}$ in $\mathcal{V}/\mathcal{X}$, and use the resulting 'residual likelihood' for purposes of estimating θ. So far as its dependence on the data is concerned, the residual likelihood is a function on the space $\mathcal{V}/\mathcal{X}$, i.e. constant on the cosets of $\mathcal{X}$.

0.10.2 *Annihilator of a subspace*

Let $\mathcal{V}$ be a vector space of dimension n, $\mathcal{X}$ a subspace of $\mathcal{V}$ of dimension p, and let $\mathcal{V}^*$ be the dual space of linear functionals on $\mathcal{V}$. Let $\mathcal{X}^0$ be a that subset of linear functionals taking the value zero on $\mathcal{X}$. Then $\mathcal{X}^0$ is a subspace of $\mathcal{V}^*$ of dimension $n - p$, called the annihilator of $\mathcal{X}$ (Halmos, 1958, section 17).

The elements of $\mathcal{X}^0$ are linear functionals taking the value zero on $\mathcal{X}$, and therefore constant on the cosets of $\mathcal{X}$. To say the same thing in another way, $\mathcal{X}^0$ is the dual space of linear functionals on the quotient space $\mathcal{V}/\mathcal{X}$, and we may write formally $\mathcal{X}^0 = (\mathcal{V}/\mathcal{X})^*$. A basis in $\mathcal{X}^0$ thus serves as the component vector of a 'point' in the quotient space relative to the dual basis.

0.11 Determinants and measures

0.11.1 *Measures and densities*

A measure μ in $\mathcal{V}$ is a real-valued function defined on a collection of subsets of $\mathcal{V}$. The main property of a measure is that it is additive for disjoint sets: $\mu(A \cup B) = \mu(A) + \mu(B)$ if A and B are disjoint. A measure thus encompasses a variety of notions such as length, area, volume, electric charge, mass and probability, all of which are additive for disjoint sets. Some measures, such as counting measure on integers, are purely atomic in the sense that μ is concentrated on a countable number of points. Other measures are continuous or a mixture of atomic and continuous parts. In the majority of applications, measures are non-negative. However, signed measures do occur, for example in electrostatics where $\mu(A)$, the net charge in A, may be positive or negative.

While it is possible, and conceptually helpful, to conceive of a measure, or any other function on $\mathcal{V}$, in the abstract, most computations require functions to be expressed in terms of the component vector x^i relative to a given basis or coordinate system. In what follows, we consider f to be a function on $\mathcal{V}$, not directly defined in a coordinate-free manner, but indirectly as a function of the component-vector x^i in $\mathcal{R}^n$. The physical appearance of such a function thus depends on the coordinate system used. Typically, therefore, f_x and f_y represent the same function on $\mathcal{V}$, but with respect to different coordinate systems.

Continuous measures have the property that μ can be expressed as an integral with respect to Lebesgue measure in the form

$$\mu(S) = \int_S f_x(x)\, dx,$$

where f_x is called the density function of the measure relative to the given basis vectors. This is an integral with respect to Lebesgue measure in the space of component vectors. However, the subscript on the integral is a subset of $\mathcal{V}$, not a subset of the component vectors in $\mathcal{R}^n$. To make sense of this, we agree that the symbol $\int_S$ denotes an integral over those component vectors x^i such that $x^i e_i$ lies in S.

One awkwardness of this construction in the present context is the dependence on the units of measurement and on the coordinate system. The Lebesgue measure of the set $[0,1]^n$ in $\mathcal{R}^n$, or the set of component vectors x^i with $0 \le x^i \le 1$ is the dimensionless number one. The meaning of this number clearly depends on whether the basis vectors e_i are in units of centimetres or kilometres, and indeed, on whether they are mutually perpendicular. The interpretation given to a unit interval in the first coordinate direction is the magnitude of e_1, the first basis vector. This unit of measure is denoted by $\|e_1\| = \langle e_1,\, e_1\rangle^{1/2}$. Assuming that the basis vectors have the physical dimension of length, the interpretation of the unit square $0 \le x^1, x^2 \le 1$ in the plane of the first two components is that of area. The units of area are given by the square root of the determinant of the 2×2 matrix with components $g_{ij} = \langle e_i,\, e_j\rangle$. For example, if e_1 denotes 'metres east' and e_2 denotes 'centimetres vertical', then $g_{11} = 10^6$, $g_{12} = g_{21} = 0$, and $g_{22} = 1$, the fundamental unit of measure being cm^2. Then the unit interval $0 \le x^1, x^2 \le 1$ in the space of component vectors represents an area of $1 \times |G|^{1/2}\ \mathrm{cm}^2$. This notion extends without modification to spaces of higher dimension.

In general, therefore, since a measure has physical units inherited from the magnitudes and orientations of the basis vectors, it is preferable to represent the density in the form

$$\mu(S) = \int_S p(x)\, |G|^{1/2}\, dx. \qquad (*)$$

The introduction of the determinantal factor may seem inconsequential because G is a constant matrix and the determinant is a constant factor in the integral. However, because G is a covariant tensor of order two, the latter representation is invariant with respect to coordinate transformation, as we shall show presently. In fact, we could replace G in $(*)$ by any symmetric positive definite covariant tensor because this implies a simple multiplicative re-definition of the units of measure.

Consider now the effect of a change of basis vectors from the original e_j to $\bar{e}_i = a_i^j e_j$ with matrix A. The component vectors tranform contravariantly from x^i to $y^i = b_j^i x^j$ with matrix $B = A^{-1}$. The transformed metric tensor is $\bar{g}_{rs} = a_r^i a_s^j g_{ij}$, so the transformed units of measure are $|\bar{G}|^{1/2} = |G|^{1/2} \times J$, where J is the absolute value of the determinant of A. Making the required change of variables $(y = Bx)$ in the integral $(*)$, we find

$$\mu(S) = \int_S p(Ay)\, |\bar{G}|^{1/2}\, dy = \int_S \bar{p}(y)\, |\bar{G}|^{1/2}\, dy.$$

The subscript on the integral is unchanged because of our convention that this is an integral over those component vectors, y, such that $y^i \bar{e}_i$ lies in $S \subset \mathcal{V}$.

From the point of view of tensor transformation properties, any function f_x that undergoes transformation by the rule

$$f_y(y) = f_x(Ay)\, J$$

when the component vectors are transformed linearly by $y = Bx$, is called a density on $\mathcal{V}$. The term *relative invariant of weight one* is also used for the same purpose. By this definition, the square root of the determinant of a second-order covariant tensor is a density. Also, the inverse square root of the determinant of a second-order contravariant tensor is a density.

By way of contrast, a function that transforms by the rule $f_y(y) = f_x(Ay)$ is called an invariant, or an absolute invariant on $\mathcal{V}$. The function $p(x)$ in $(*)$ is an absolute invariant. A likelihood function, unlike a probability density function, is an absolute invariant in this sense, with respect to re-parameterization. Potential functions associated with gravitational or electrostatic fields are examples of absolute invariants in physics.

0.11.2 *Factors of measures*

Let $\mathcal{V}_1, \mathcal{V}_2$ be complementary subspaces of $\mathcal{V}$ such that $e_1, \ldots, e_m$ is a basis in $\mathcal{V}_1$, and $e_{m+1}, \ldots, e_n$ is a basis in $\mathcal{V}_2$. The fundamental metric tensor on $\mathcal{V}$, with components $g_{ij} = \langle e_i, e_j \rangle$, can be partitioned as follows.

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}; \qquad G^{-1} = \begin{pmatrix} G^{11} & G^{12} \\ G^{21} & G^{22} \end{pmatrix}.$$

Thus, G_{11} is the metric tensor on $\mathcal{V}_1$, and G_{22} is the metric tensor on $\mathcal{V}_2$. Further, $|G|^{1/2}$ is a density on $\mathcal{V}$, $|G_{11}|^{1/2}$ is a density on $\mathcal{V}_1$, and $|G_{22}|^{1/2}$ is a density on $\mathcal{V}_2$.

Now, G^{-1} is the fundamental metric tensor on $\mathcal{V}^*$, so G^{22} is the metric tensor on $\mathcal{V}_2^*$, the dual space of linear functionals on $\mathcal{V}_2$. The dual basis is the set of linear functionals v^j such that $v^j(e_i) = \delta_i^j$. Consequently the basis linear functionals in $\mathcal{V}_2^*$ all take the value zero on $\mathcal{V}_1$, so $\mathcal{V}_2^*$ is the annihilator of $\mathcal{V}_1$. Also, $\mathcal{V}_2^*$ is the space of linear functionals on the quotient space $\mathcal{V}/\mathcal{V}_1$. Conversely, $\mathcal{V}_1^*$ is the annihilator of $\mathcal{V}_2$, and the space of linear functionals on $\mathcal{V}/\mathcal{V}_2$.

Using formulae for the inverse of a partitioned matrix, G^{22} satisfies

$$G^{22} = \left(G_{22} - G_{21} G_{11}^{-1} G_{12}\right)^{-1}.$$

The determinant of G is given by

$$\begin{aligned} |G|^{1/2} &= |G_{11}|^{1/2}\, |(G_{22} - G_{21} G_{11}^{-1} G_{12})|^{1/2} \\ &= |G_{11}|^{1/2}\, |G^{22}|^{-1/2}. \end{aligned} \qquad (**)$$

In words, the density $|G|^{1/2}$ on $\mathcal{V}$ factors into a product of two densities, $|G_{11}|^{1/2}$ on $\mathcal{V}_1$, and $|G^{22}|^{-1/2}$ on $\mathcal{V}/\mathcal{V}_1$. The alternative factorization $|G^{11}|^{-1/2}|G_{22}|^{1/2}$ is also a product of two densities, on $\mathcal{V}/\mathcal{V}_2$ and $\mathcal{V}_2$, showing that the factorization is not unique.

To understand better the nature of these two factorizations, consider a change of basis vectors in $\mathcal{V}$ in which the subspace $\mathcal{V}_1$ is preserved. In other words, consider a re-parameterization $\bar{e}_r = a_r^i e_i$ in which a_r^i is zero for $r \le m$ and $i > m$. In matrix notation, if r denotes the row and i the column, the components a_r^i and b_r^i of A and B are

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}; \qquad B = A^{-1} = \begin{pmatrix} A^{11} & 0 \\ A^{21} & A^{22} \end{pmatrix}.$$

Note that A^{11} is the inverse matrix of A_{11}, and likewise for A^{22}. The transformed metric tensor and its inverse are AGA^T and $B^T G^{-1} B$, or, in tensor notation, $a_r^i a_s^j g_{ij}$ and $b_r^i b_s^j g^{rs}$. The first factor in $(**)$,

$$\bar{G}_{11} = A_{11}^T G_{11} A_{11}$$

is the transformed metric tensor on $\mathcal{V}_1$. The complementary factor

$$(\bar{G}^{22})^{-1} = A_{22}^T(G_{22} - G_{21}G_{11}^{-1}G_{12})A_{22}$$

functions as a metric tensor on any space complementary to $\mathcal{V}_1$, i.e. on the space $\mathcal{V}/\mathcal{V}_1$. The components of this tensor are unaffected by the sub-matrix A_{21} in the transformation.

0.11.3 *Singular normal likelihood*

Let Y be a normally distributed random variable in $\mathcal{V} = \mathcal{R}^n$ with mean μ, and covariance matrix Σ depending on unknown parameters to be estimated. It is convenient to let $\mathcal{V}$ be an inner product space with inner product matrix $G = \Sigma^{-1}$. The joint density of Y at y is given by

$$f(y) = (2\pi)^{-n/2}\,|G|^{1/2}\,\exp\bigl(-\tfrac{1}{2}\|y-\mu\|^2\bigr),$$

where $\|y\|^2 = \langle y,\, y\rangle = y^TGy$ is the invariant squared length of y.

Suppose that $\mu = E(Y)$ is known to lie in a given p-dimensional subspace $\mathcal{X}$, specified, for example, as the span of the columns of a matrix X. Let $\bar{e}_1, \ldots, \bar{e}_p$ be a basis in $\mathcal{X}$, and let $\bar{e}_{p+1}, \ldots, \bar{e}_n$ be a basis in any complementary space such that the transformation from the old to the new basis has unit determinant. Let P be the orthogonal projection from $\mathcal{V}$ to $\mathcal{X}$, and let $Q = I - P$ be the complementary projection. Note in particular that QY is a residual vector in $\mathcal{V}/\mathcal{X}$. Pythagoras's theorem gives $\|y\|^2 = \|Py\|^2 + \|Qy\|^2$. Further, the transformed inner product matrix with respect to the new basis is

$$\bar{G} = \begin{pmatrix} \bar{G}_{11} & \bar{G}_{12} \\ \bar{G}_{21} & \bar{G}_{22} \end{pmatrix}.$$

The determinant of $\bar{G}$ factors as the product

$$|G| = |\bar{G}| = |\bar{G}_{11}|\,|\bar{G}^{22}|^{-1}, \qquad (*)$$

such that $|\bar{G}_{11}|^{1/2}$ is a density on $\mathcal{X}$, and $|\bar{G}^{22}|^{-1/2}$ is a density on $\mathcal{V}/\mathcal{X}$. Then the joint density of Y factors into a product as follows:

$$(2\pi)^{-n/2}|\bar{G}_{11}|^{1/2}\exp\bigl(-\tfrac{1}{2}\|P(y-\mu)\|^2\bigr) \times |\bar{G}^{22}|^{-1/2}\exp\bigl(-\tfrac{1}{2}\|Qy\|^2\bigr).$$

The first factor is the density of the fitted values in $\mathcal{X}$. The second is the residual density on $\mathcal{V}/\mathcal{X}$, constant on cosets of $\mathcal{X}$.

The matrix $\bar{G}_{11}$ is the induced metric tensor on $\mathcal{X}$. In matrix notation, if the columns of X are used as the basis in $\mathcal{X}$, this is X^TGX, or $X^T\Sigma^{-1}X$. From equation (*), we have

$$|\bar{G}^{22}|^{-1} = |G|/|\bar{G}_{11}| = |\Sigma|^{-1}\,|X^T\Sigma^{-1}X|^{-1}.$$

Consequently, the likelihood function based on the singular normal distribution of the residuals in $\mathcal{V}/\mathcal{X}$ is

$$|\Sigma|^{-1/2}\,|X^T\Sigma^{-1}X|^{-1/2}\exp\bigl(-\tfrac{1}{2}\|Qy\|^2\bigr) \qquad (**)$$

In this expression, Q is the orthogonal complement of P, not an arbitrary projection having null space $\mathcal{X}$. As a matrix Q is given by $I - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$.

This derivation asumes that the transformation to the new basis vectors spanning $\mathcal{X}$ and its complement has unit determinant. If this assumption is not satisfied, a determinantal factor independent of Σ enters into the density calculations. But this determinant is a constant and has no effect on the residual likelihood, so (**) is unaffected.

0.12 Vector Fields

0.12.1 *Background*

A vector field on a set S is a vector-valued function on S. To each element $x \in S$ is assigned a vector $f(x)$, read as 'f at x.' In fluid dynamics, for example, x denotes position in real space, and f at x is the velocity vector of the fluid at x. In gravitation, mechanics or electrostatics, $f(x)$ may denote the force field at x. In applications connected with statistical estimation, x is a point in the parameter space, and $f(x)$ is the value of the estimating function at x. In most physical applications, it is appropriate to imagine the collection of vectors at different points in S as vectors in the same space. Such vectors can therefore be added or subtracted. For example, the average velocity of a fluid is obtained by adding the velocity vectors of all the particles and dividing by the number of particles. But, in general, although $f(x_1)$ may be numerically equal to $f(x_2)$, the physical effects (on the motion of a rigid body, for example), may be quite different. In statistical applications, the metric tensor or covariance matrix frequently depends on the parameter. Even if $f(x_1)$ is numerically equal to $f(x_2)$, an assessment of statistical significance may depend on the metric tensor, leading to a difference in conclusions. In general, therefore, $f(x)$ is best regarded as a vector in a space $\mathcal{T}_x$, usually the tangent space at x or its dual. Each point has its own tangent space, possibly with its own metric tensor. The collection of these spaces is called the tangent bundle.

Physical conservation laws require fields to have certain characteristic properties. For example, if f is an electrostatic or gravitational force field, conservation of energy implies that the net work required or energy gained in traversing a closed loop γ must be zero. Otherwise, energy could be created by traversing the loop indefinitely. Now, f is a vector in the same space as x, or in the tangent space, so it is appropriate to denote the components by x^i and f^i, in the manner of contravariant tensors. The conservation of energy condition is then

$$\int_\gamma \langle f,\, dx\rangle = \int_\gamma f^i\, g_{ij}\, dx^j = 0,$$

where g_{ij} is the metric tensor associated with the inner product. Fields satisfying this condition are called conservative. Likewise, conservation of mass considerations in the flow of incompressible fluids require that, for any element of volume, the net inflow should equal the net outflow, a condition known as zero divergence.

0.12.2 *Divergence of a vector field*

Let $\mathcal{V}$ be an affine space, and let $f(\cdot)$ be a vector field defined on $\mathcal{V}$ with values in $\mathcal{T}$, the tangent space of $\mathcal{V}$. The components of f at v are denoted by $f^i(v)$. Many familiar examples of vector fields in physics are of this type. Real space has no natural origin, so it is best envisaged as an affine space (ignoring relativistic curvature due to gravitation). Although x is a point in $\mathcal{V}$, f itself is a vector in the tangent space of $\mathcal{V}$. Now, $\partial/\partial x^i$ are the components of a vector in $\mathcal{T}^*$, the dual of $\mathcal{T}$. Consequently, assuming that $f(\cdot)$ is differentiable,

$$f^i_j(x) = \partial f^i(x)/\partial x^j$$

are the components at $x \in \mathcal{V}$ of a tensor in $\mathcal{T} \otimes \mathcal{T}^*$. Further, since δ^j_i are the components of a tensor in the same space, the scalar $d(x) = \delta^j_i f^i_j(x)$ is an invariant, known as the divergence of the field $f(\cdot)$ at x.

Invariance of the divergence of a contravariant vector field is an automatic consequence of tensor notation, but tensor notation does not offer an explanation of what divergence measures. In fact, divergence of f at x_0 is the average value of $\langle f,\, x - x_0\rangle/\langle x - x_0,\, x - x_0\rangle$. This average must be interpreted in an invariant sense as an integral with respect to invariant measure over the surface

or interior of the infinitesimal sphere $\langle x - x_0, x - x_0\rangle \leq \rho^2$ centered at x_0. In this sense, $\rho d(x_0)$ is the surface-average radial component of the field on the surface of this sphere: $d(x_0)$ is the volume-average of $\langle f, x - x_0\rangle / \langle x - x_0, x - x_0\rangle$ over the interior of the sphere. In fluid mechanics, the divergence of the velocity vector at x is proportional to net the outward flow of fluid from x. Incompressible fluids are characterized by zero divergence at every point. Fields that have zero divergence are said to be solenoidal, Aris (1989, section 3.43).

0.12.3 *Rotational fields*

A rotational field is a mathematical model of vortex motion. The simplest rotational field is one in which the field is linear in x and orthogonal to x at all points. Now, $\langle x, Ax\rangle = x^T GAx$ is identically zero if and only if GA is skew-symmetric. Thus, a necessary and sufficient condition that the field $y^r = a^r_i x^i$ be purely rotational is that $a_{ik} = g_{ij}a^j_k$ be a skew-symmetric covariant tensor. More generally, a non-linear contravariant vector field f is said to be irrotational at x if its partial derivatives $f^i_r = \partial f^i / \partial x^r$ satisfy the symmetry condition $g_{ik}f^k_j(x) = g_{jk}f^k_i(x)$. A field that is irrotational at all points is said to be conservative.

In general, rotational motion and turning forces, such as angular velocity, angular momentum, magnetic fields and torque, are represented in physics by skew-symmetric matrices. For example, the positive, or anticlockwise, moment of the force (f^1, f^2) applied at the point (x^1, x^2) in the plane is equal to $x^1 f^2 - x^2 f^1$. More generally, the moment of a force is represented by the skew-symmetric contravariant tensor $x^i f^j - x^j f^i$. This is simply an alternative, and in many ways preferable, representation of the cross product or vector product $x \times f$ in $\mathcal{R}^3$.

0.12.4 *Curl of a covariant vector field*

Let f be a covariant vector field on an affine space $\mathcal{V}$. In statistics, a score function or vector estimating function is of this type. The components of f in $\mathcal{T}^*$, the dual of the tangent space, are denoted by $f_r(x)$. The partial derivatives of f form a covariant tensor of order two with components $f_{rs} = \partial f_r / \partial x^s$. Any square matrix can be decomposed into a symmetric and a skew-symmetric part:

$$2f_{rs} = (f_{rs} + f_{sr}) + (f_{rs} - f_{sr}).$$

The first term represents the components of a tensor in $\mathrm{sym}(\mathcal{T}^{*\otimes 2})$: the second term gives the components of a tensor in the complementary space $\mathrm{alt}(\mathcal{T}^{*\otimes 2})$. The latter component is called the curl of the field f, generalizing the familiar definition for vectors in $\mathcal{R}^3$. For further details, see Synge and Schild (1978, p. 246).

The term curl originates in Maxwell's work in electromagnetism, but this particular combination of derivatives had arisen earlier in work on the refraction of light (MacCullagh, 1846).

0.12.5 *Newton's second law*

Let $x^i \equiv x^i(t)$ be the position at time t of a particle of mass m subject to a force with components f^i. Newton's second law states that the force is equal to the rate of change of momentum. In symbols,

$$f^i = \frac{d}{dt}(mv^i) = \frac{d}{dt}(m\dot{x}^i).$$

This is a contravariant vector equation. If there are many particles in the system, we may sum these vector contributions, one vector for each particle. Then the total vector force is equal to the rate of change of the total momentum vector.

The motion of a rigid body is constrained to translation and rotation. The following equations, derived from Newton's second law, relate directly to rotary motion. The moment about the origin

of a force f applied at x is given by the skew-symmetric tensor $M^{rs} = x^r f^s - x^s f^r$. The angular momentum, or moment of momentum, is given by

$$H^{rs} = m(x^r v^s - x^s v^r),$$

where v denotes the velocity vector. Differentiation of H, followed by application of Newton's second law, gives $dH^{rs}/dt = M^{rs}$. In words, the moment of the force on a particle is equal to the rate of change of angular momentum.

Further details of the application of tensor calculus to classical dynamics can be found in Synge and Schild (1978, chapter 5).

0.13 Invariants

An invariant is a function whose value is unaffected by a relevant class of transformations. The class of transformations is usually implied by the context. The principal idea is that fundamentally important aspects of a problem are those that are unaffected by unimportant or cosmetic details such as the choice of basis vectors or the choice of coordinate system in parametric statistics. Tensor notation, although paradoxically very much tied to coordinate systems, does generate invariants in an automatic fashion as tensors of order zero. It does not supply a geometric interpretation of such invariants.

The invariance argument is mathematically very powerful, and quite persuasive in many circumstances. But it can be carried to excess, particularly in problems where the statistical model has unintended or accidental symmetries that the physical process does not possess. For example, the multivariate normal model has rotational symmetries that physical processes usually do not possess. Likewise, the univariate Cauchy model has symmetries beyond the usual location and scale (Exercise ?.?).

To give a simple example of an invariant, let ω be a linear transformation from a vector space to itself. In terms of component vectors, the transformation may be written in the form

$$y^r = \omega_i^r x^i,$$

and ω_i^r are the components of a mixed tensor in $\mathcal{V} \otimes \mathcal{V}^*$. Now, δ_i^r are the components of the identity function, a mixed tensor in the same space. Since the tensor property is preserved under multiplication and contraction, it follows that

$$\lambda = \omega_i^r \delta_r^i$$

is a tensor of order zero, or an invariant. In matrix notation, a change of basis in $\mathcal{V}$ induces a tensorial transformation from the original Ω to a new matrix of the form $\bar{\Omega} = A^{-1}\Omega A$. This is known as a similarity transform or unitary transform (Dirac, 1958, ch. 26). In Cartan's terminology (Cartan, 1981, section 41), the matrices Ω and $\bar{\Omega}$ are said to be equivalent. It is shown in Exercise 1.? that the eigenvalues of a (1,1) tensor are invariant under this transformation. In other words, equivalent matrices have the same eigenvalues. In particular, $\lambda = \mathrm{tr}(\Omega)$ is equal to the sum of the eigenvalues and is invariant.

The second example is closer in spirit to the material in the following chapters in that it involves random variables in an explicit way. Let $\sigma^{i,j}$ be the covariance matrix of X, and consider the effect of making an *orthogonal* transformation from X to new variables $\bar{X} = AX$. In matrix notation Σ is transformed to $\bar{\Sigma} = A\Sigma A^T$. Since the characteristic polynomial of $\bar{\Sigma}$ is the same as the characteristic polynomial of Σ, it follows that the eigenvalues are invariant. This invariance holds only for orthogonal transformation of X. Sylvester's law of inertia states that only the rank and signature are invariant under linear transformation. Reverting to index notation, $\sigma^{i,j}\delta_{ij}$, the trace of Σ, is invariant under orthogonal transformation because both $\sigma^{i,j}$ and δ_{ij} transform as tensors

under this group. δ_{ij} is not a tensor under the general linear group, so the indicated trace is not invariant under the larger group.

Log likelihood functions are automatically invariant under invertible transformation of the sample space. When we talk of invariance in this context, we usually mean invariance under a group of transformations on the parameter space. The log likelihood itself is invariant under re-parameterization, but its derivatives are not. The log likelihood ratio statistic $l(\hat{\theta}; y) - l(\theta; y)$ is invariant, and consequently its distribution is invariant. Chapter ? discusses ways of approximating the distribution in terms of invariants.

Example 27: Mahalanobis distance. Let $X = (X^1, \ldots, X^n)$ be the components of a random vector in an affine space $\mathcal{V}$ with tangent space $\mathcal{T}$. Example 10 shows that the covariance matrix $\kappa^{i,j}$ are the components of a tensor in $\mathcal{T}^{\otimes 2}$, and that the inverse covariance matrix $\kappa_{i,j}$ are the components of a tensor in $\mathcal{T}^* \otimes \mathcal{T}^*$. In other words, $\kappa_{i,j}$ is a metric tensor on $\mathcal{V}$. If X_1 and X_2 are two points in $\mathcal{V}$, then $d = X_1 - X_2$ lies in $\mathcal{T}$, and $d^i d^j$ are the components of a tensor in $\mathcal{T}^{\otimes 2}$. Consequently, the Mahalanobis squared distance $d^i d^j \kappa_{i,j}$ is invariant under affine transformation in $\mathcal{V}$.

0.14 Densities and invariant measures

A measure μ in a set $\mathcal{V}$ is said to be invariant under the group $\mathcal{G}$ acting on $\mathcal{V}$ if, for each $A \subset \mathcal{V}$ and $g \in \mathcal{G}$,

$$\mu(gA) = \mu(A)$$

where gA denotes the translation of A by g.

An invariant measure need not exist: if it exists it need not be unique. The set of translates of $x \in \mathcal{V}$ by $\mathcal{G}$ is called an *orbit*, and x is an *orbit representative*. If $\mathcal{V}$ comprises a single orbit in the sense that $\mathcal{V} = \{gx : g \in \mathcal{G}\}$, then the action is said to be *transitive*. If the action is transitive and if an invariant measure exists, then it is unique up to a scalar multiplier.

Three examples will illustrate the ideas. Let $\mathcal{V}$ be the finite set $\{1, \ldots, n\}$, and let $\mathcal{G}$ be the symmetric group of permutations. The action is transitive, and cardinality measure or counting measure on subsets is invariant. However, if $\mathcal{G}$ is the group of cyclic permutations, $\mathcal{V}$ comprises $(n-1)!$ orbits, and the action is not transitive. In this case, any measure constant on orbits is invariant. Second, Lebesgue measure in $\mathcal{R}$ is invariant under translation: there is no measure that is invariant under location-scale transformation on $\mathcal{R}$. A less obvious example is the following. Let $\mathcal{V}$ be the complex plane (omitting the real axis and with conjugate pairs identified). Let $\mathcal{G}$ be the group of fractional linear transformations on $\mathcal{V}$ of the form

$$gz = \frac{az + b}{cz + d},$$

with a, b, c, d real and $ad - bc \neq 0$. It should be checked that this is indeed a group, and that the 2×2 matrix with elements a, b, c, d is a representation of g, group composition corresponding to matrix multiplication. Invariant measure on subsets of the upper half plane, ($z = x + iy$ with $y > 0$), is given by

$$\mu(A) = \int_A dx\, dy / y^2$$

provided that this integral exists. It is not obvious, but it is nonetheless true, that $\mu(gA) = \mu(A)$. This measure is also invariant under all subgroups of $\mathcal{G}$, in particular the location-scale subgroup $(a, b) \circ (x, y) \mapsto (x + a, by)$, with $b > 0$, acting on the upper half plane.

0.15 Bibliographic notes

For an excellent orthodox and comprehensive introduction to finite-dimensional vector spaces, see Halmos (1958). For a more iconoclastic description in line with sections 1 and 2 of this chapter, see Hoffmann (1966).

We have chosen in this chapter to include all units of measurement in the definition of the basis vectors. As a consequence, component vectors are ordered sets of real numbers. This practice is not universal: for a forcefully argued contrary view in which symbols denote physical magnitudes, see Jeffreys and Jeffreys (1956, p. 3–4). In most treatments, particularly in undergraduate physics texts, it is unclear where the units of measurement reside. Kibble (1985, p. 4) gives an abbreviated discussion but his convention is not entirely clear on this point. Most mainstream mathematics texts such as Halmos (1958) or Birkoff and MacLane (1948) seem to regard units of measurement as a non-issue not worth discussing. Other than in examples, the basis vectors are never *defined* in physical terms, so this hands-off approach is mathematically quite defensible, perhaps even necessary. Hoffmann (1966, p. 57), however, is explicit in using the same convention as this book. Among statistics texts, the conventions are often not spelled out; an exception is Stone (1986) whose convention coincides with ours.

Tensor notation is widely used in applied mathematics, to areas such as fluid dynamics, elasticity, electromagnetism, relativity, and differential geometry. Definitions and notation vary to some extent from subject to subject and from author to author. For example, Jeffreys (1952) and Jeffreys and Jeffreys (1956) are concerned with the effect on equations of motion of rotating the axes or frame of reference. Consequently, their definition of what they call a Cartesian tensor refers only to the group of rigid motions (rotation and translation), and not, in general, to reflections or linear transformations.

Other useful references, again with a bias towards applications in physics, include, in order of first publication, McConnell (1931), Synge and Schild (1978), Schouten (1989), Aris (1989), Wrede (1972), Thomas (1965), Lawden (1968), and Richtmyer (1981). Most of these authors define a tensor as a set of quantities that obey the tensor transformation rules under a change of coordinates. With the exception of Richtmyer, there is little discussion of vector spaces and dual spaces.

In connection with differential geometry and calculus on manifolds, see the books by Eisenhart (1926), Weatherburn (1950), Sokolnikoff (1951), Stoker (1969), Lovelock and Rund (1989), Bishop and Goldberg (1980), Willmore (1982) and Spivak (1970).

The summation convention was first introduced by Einstein (1916) in his paper on the foundation of general relativity.

For a discussion of the geometry of generalized inverses, see Kruskal (1975), Rao (1973), or Stone (1987).

0.16 Further results and exercises 1

0.1 Vectors in $\mathcal{R}^2$ are sometimes specified in polar coordinates by magnitude and direction. All vectors have a definite magnitude. All except one have a definite direction. Which one?

0.2 In a non-metric space the notion of length or distance is absent. Bearing this in mind, answer question 7 of section 0.1.1.

0.3 Show that the 'parallelogram law of addition' for component vectors is a consequence of the definition of linear functionals in section 0.3.1.

0.4 Let $\{1, x, x^2\}$ be a basis for quadratic functions on $[0, 1]$. What are the components of $v(x) = 2x^2 - 3x + 1$ with respect to this basis? What are the components of v with respect to the basis $\{1, 2x - 1, 6x^2 - 6x + 1\}$?

0.5 For square-integrable functions on $[0, 1]$, define the inner product

$$\langle v_1,\, v_2\rangle = \int_0^1 v_1(x)\, v_2(x)\, dx.$$

Show that for the first choice of basis in the preceding exercise,

$$G = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}, \qquad G^{-1} = \begin{pmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{pmatrix}.$$

Use equation (1.2) to find the components of $v(x) = x^2 - 3x - 1$ with respect to the two bases given in the preceding exercise.

0.6 Let $\mathcal{V}$ be the extended complex plane, and let the modulus and argument of v_j be denoted by $|v_j|$ and ϕ_j respectively. The point zero is defined to have argument zero. Multiplication by real scalars is defined in the standard way, but addition is defined as follows.

$$v_1 + v_2 = \frac{|v_1| + |v_2|}{1 - |v_1|\,|v_2|} e^{i\phi_1 + i\phi_2}.$$

Show that the associativity condition $v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3$ is satisfied. Examine the conditions in section 2 of Halmos (1958) and explain why $\mathcal{V}$ is or is not a vector space.

0.7 Show that the eigenvalues of a real skew-symmetric matrix are purely imaginary ($\bar{\lambda} = -\lambda$). Hence deduce that a skew-symmetric matrix of odd order must have at least one zero eigenvalue.

0.8 Let $A^{ij} = a^i b^j - a^j b^i$ be a skew-symmetric matrix of order three. Show that the vector product $c = a \times b$, as defined in section 1.4.5, lies in the null space of A.

0.9 Let X be a $n \times p$ matrix of rank $q < p \leq n$, and let W be symmetric positive definite of order n. Define two matrices P_1 and P_2, both of the form

$$P = X(X^T W X)^- X^T W,$$

but with different choices of generalized inverse. Show that P is a self-conjugate projection matrix if the generalized inverse is annihilating. On the assumption that both generalized inverses are annihilating, and that the rank of $X^T W X$ is the same as the rank of X, show that $P_1 = P_2$.

0.10 Show that the identity matrix is a generalized inverse of any projection matrix. Show also that a projection matrix, not necessarily symmetric, is its own annihilating generalized inverse.

0.11 Let ω^{ij}, with inverse ω_{ij} be the components of a $p \times p$ matrix of rank p. Show that

$$\gamma^{rs,ij} = \omega^{ri}\omega^{sj} + \omega^{rj}\omega^{si},$$

regarded as a $p^2 \times p^2$ matrix with rows indexed by (r, s) and columns by (i, j), is symmetric with rank $p(p+1)/2$. Show also that $\omega_{ri}\omega_{sj}/2$ is a generalized inverse.

0.12 *Singular values of a matrix*: Let the $p \times q$ matrix A, with components a_i^r, be considered as defining a linear transformation from the domain, $\mathcal{R}^q$, to the range in $\mathcal{R}^p$. Interpret the *singular values* of A as invariants under independent orthogonal transformation of the domain and range spaces. For the definition of singular values and their application in numerical linear algebra, see Chambers (1977, Section 5.e).

0.13 Let A, A^{-1} and X be symmetric matrices with components a_{ij}, a^{ij} and x_{ij} respectively. Show that the Taylor expansion for the log determinant of $A+X$ about the origin may be written

$$\log\det(A+X) = \log\det(A) + x_{ij}a^{ij} - x_{ij}x_{kl}a^{ik}a^{jl}/2 + x_{ij}x_{kl}x_{mn}a^{ik}a^{jm}a^{ln}/3 + ...$$

Describe the form of the general term in this expansion. Generalize this expansion to asymmetric matrices.

0.14 *Hooke's Law*: In the mechanics of deformable solids, the components of the *stress tensor*, p_{ij}, measure force per unit area in the following sense. Let e_i be the unit vector in the ith coordinate direction and let $\bar{e}_i$ be the orthogonal plane. Then p_{ii} is the force per unit area normal to the plane, also called the *normal stress*, and p_{ij}, $j \neq i$ are the *shear stresses* acting in the plane. The components of the *strain tensor*, q_{ij}, which are dimensionless, measure percentage deformation or percentage change in length. Both arrays are symmetric tensors under the orthogonal group.

In the case of *elastic* deformation of an *isotropic* material, Hooke's law in its most general form states that the relationship between stress and strain is linear. Thus,

$$p_{rs} = b^{ij}_{rs}q_{ij},$$

where b^{ij}_{rs} is an isotropic fourth-order tensor given by

$$b^{ij}_{rs} = \lambda\delta^{ij}\delta_{rs} + 2\mu\delta^i_r\delta^j_s,$$

for constants λ, μ that are characteristic of the material.

Show that the inverse relationship giving the strains in terms of the stresses may be written in the form

$$q_{ij} = \left(\lambda'\delta^{rs}\delta_{ij} + 2\mu'\delta^r_i\delta^s_j\right)p_{rs},$$

where the new constants are given by

$$\mu' = \frac{1}{4\mu}, \qquad \lambda' + 2\mu' = \frac{\lambda+\mu}{\mu(3\lambda+2\mu)} = E^{-1}.$$

In the terminology used in Mechanics, E is known as *Young's modulus* or *modulus of elasticity*, μ is called the *rigidity* or *shear modulus* and $\sigma = \lambda/\{2(\lambda+\mu)\}$ is *Poisson's ratio*. Note that $E = 2(1+\sigma)\mu$, implying that two independent constants entirely determine the three-dimensional elastic properties of the material. (Murnaghan, 1951, Chapters 3,4; Jeffreys & Jeffreys 1956, Section 3.10; Drucker, 1967, Chapter 12).

0.15 *Alternating tensor*: The permutation symbol or alternating tensor of order n is defined as follows:

$$\epsilon^{i_1,\ldots,i_n}_{r_1,\ldots,r_n} = \begin{cases} 1 & \text{if } r_1,\ldots,r_n \text{ is an even permutation of } i_1,\ldots,i_n; \\ -1 & \text{if } r_1,\ldots,r_n \text{ is an odd permutation of } i_1,\ldots,i_n; \\ 0 & \text{otherwise.} \end{cases}$$

The value zero is understood to apply if any subscript or any superscript is duplicated. Show that, for any $n \times n$ matrix A

$$\epsilon^{i_1,\ldots,i_n}_{1,\ldots,n}a^1_{i_1}a^2_{i_2}\cdots a^n_{i_n} = \det(A).$$

Let $\mathcal{V}$ be a vector space of dimension n. Let ϵ be the component matrix of a transformation from $\mathcal{V}^{\otimes n}$ to itself as follows:

$$y^{i_1,\ldots,i_n} = \epsilon^{i_1,\ldots,i_n}_{r_1,\ldots,r_n}x^{r_1,\ldots,r_n}.$$

where x and y are tensors in $\mathcal{V}^{\otimes n}$. Now introduce a change of basis vectors in $\mathcal{V}$ as follows: $\bar{e}_r = a^i_r e_i$. Prove that the component vectors in the new coordinate system satisfy

$$\bar{y}^{i_1,\ldots,i_n} = \epsilon^{i_1,\ldots,i_n}_{r_1,\ldots,r_n}\bar{x}^{r_1,\ldots,r_n},$$

showing that the components of ϵ are the same in all coordinate systems, (Synge and Schild, Ch. 7).

0.16 Define an alternative permutation symbol by

$$\epsilon_{r_1,\ldots,r_n} = \epsilon^{1,\ldots,n}_{r_1,\ldots,r_n}.$$

Show that this symbol transforms as an isotropic tensor under O^+, the orthogonal group with positive determinant (Ames and Murnaghan 1929, p. 440), (Wrede 1972, p. 85), (Jeffreys and Jeffreys, 1956, Sections 2.07, 3.03).

0.17 Explain why the multinomial generalized inverse matrix given by Watson (1996) is not the same as that given at the end of section 1.8.

0.18 *Rigid body dynamics*: Let $y(x;t)$ be the position at time t of the point originally at x at time $t = 0$. Thus, $y(x;0) = x$. For each t, $y(\cdot;t)$ is a transformation of an affine space $\mathcal{V}$ to itself satisfying the rigid-body condition $|y(x_1;t) - y(x_2;t)|^2 = |x_1 - x_2|^2$. Deduce that $y(x;t) = b(t) + A(t)x$, where $b(t) \in \mathcal{V}$, and $A(t)$ is an orthogonal matrix.

Let $v = v(x;t) = \dot{y}(x;t)$ be the velocity vector of x. Show that, for any pair of points, x_1, x_2, the velocity vectors satisfy the orthogonality condition

$$\langle v_1 - v_2,\, y_1 - y_2 \rangle = 0.$$

One solution to this differential equation is $v_1(t) = v_2(t)$, implying that all points move in parallel and in unison. Suppose, however, that the motion is such that at time t_0, one point, $y(x_0;t_0)$, is stationary, i.e. $\dot{y}(x_0;t_0) = 0$. Show then that for each $y = y(x;t_0)$,

$$\langle \dot{A}A^{-1}(y - y_0),\, y - y_0 \rangle = 0,$$

and hence that $G\dot{A}A^{-1}$ is skew-symmetric. Deduce that the covariant components of the velocity vector at time t_0 satisfy

$$v_i = g_{ij}v^j = \omega_{ij}(y^j - y^j_0),$$

where ω_{ir} is a skew-symmetric covariant tensor.