A Geometric Approach to Determinants

John Hannah

We are all happy to use pictures when we first introduce students to calculus. Why not take the same approach in linear algebra? While good progress has been made in this direction in recent years, determinants seem to have escaped this trend. Most textbooks still introduce them via cofactor expansions (see **[FB]** and **[N]** for example), the permutation definition (**[AK]**, **[Se]**), or via their alternating multilinear form properties (**[DL]**, **[St]**).

In this article I propose a geometric introduction to determinants. The details are not new, though they are well scattered through the literature. For example, a geometric view of the 2×2 case is used as motivation for an algebraic approach in **[DL]** and **[O]**. What perhaps is new (or at any rate, has not been fashionable for at least a couple of generations) is that I am suggesting that the geometric view be given a defining role similar to that given the "area under the graph" definition of the integral, which we routinely use in beginning calculus courses.

Before you push the panic button, I'm not suggesting that rigorous algebraic approaches be abandoned. What I am saying is that, particularly in the case of determinants, this approach is not very suitable for students who are meeting linear algebra for the first time. In fact, many textbooks implicitly recognize this problem by relegating some key proofs to later sections or appendices, so that students may avoid them or, at any rate, take them on trust (see **[FB]**, **[N]**). In my own institution, I see a geometric approach as being appropriate for our first year linear algebra students, while an algebraic approach is more appropriate for our advanced courses.

Just as in the calculus context, geometry helps students to form mental images or constructs that they can use to help them understand what determinants are all about. I have anecdotal evidence that it encourages students to engage in what Blum and Kirsch call "preformal" proving **[BK]**. In other words, students can see or conjecture properties of the determinant, along with (geometric) explanations appropriate to their level of mathematical development.

Another reason for this approach is that I want a treatment that focuses on the important properties of the determinant, without getting involved in issues that I see as being peripheral for first year students, most of whom will not major in mathematics. Despite this, I like to think that my shopping list will keep most mathematicians happy, too. Here is my list of "what every first year student should know about determinants"

1. $\det A$ gives the area or volume magnification factor for the linear transformation $\mathbf{x} \to A\mathbf{x}$;
2. A is invertible if and only if $\det A \neq 0$;
3. $\det(AB) = (\det A)(\det B)$;
4. the most efficient way of calculating $\det A$ is by row (or column) reducing A to triangular form.

In what follows I have sketched out the main features of this approach to determinants. I have tried to retain some of the flavour of the way I present it to students, but of course I can leave out some of the details when I am talking to you.

DEFINING THE DETERMINANT. Let A be a $n \times n$ real matrix. We can view A as a linear transformation from R^n to R^n given by $\mathbf{x} \to A\mathbf{x}$.

If A is a 2×2 matrix with column vectors $\mathbf{a}$ and $\mathbf{b}$, then the linearity means that A transforms the unit square in R^2 into the parallelogram in R^2 determined by $\mathbf{a}$ and $\mathbf{b}$ (see Figure 1).

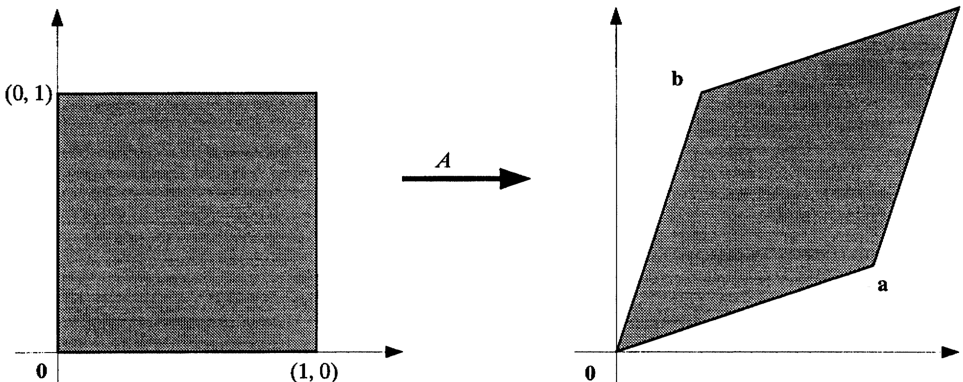

Figure 1. Effect of the matrix $A = (\mathbf{a}|\mathbf{b})$ on the unit square.

Similarly in the 3×3 case, A maps the unit cube in R^3 into the parallelepiped (or solid parallelogram) in R^3 determined by the column vectors of A. In general, an $n \times n$ matrix A maps the unit n-cube in R^n into the n-dimensional parallelogram determined by the column vectors of A.

Other squares (or cubes, or hypercubes, etc.) are transformed in much the same way and scaling the sides of the squares merely scales the sides of the parallelograms (or parallelepipeds, or higher dimensional parallelograms) by the same amount. In particular, the magnification factor

$$\frac{\text{area (or volume) of image region}}{\text{area (or volume) of original region}}$$

is always the same, no matter which squares (or cubes, or hypercubes) we start with. Indeed, since we can calculate the areas (or volumes) of reasonably nice regions by covering them with little squares (or cubes) and taking limits, the above ratio will still be the same for these regions, too.

Definition. The determinant of the matrix A is the above magnification factor.

For example, since the unit square has area 1, the determinant of a 2×2 matrix A is the area of the parallelogram determined by the columns of A. Similarly, the determinant of a 3×3 matrix A is the volume of the parallelepiped determined by the columns of A. Notice that this way of defining the determinant does not give us an obvious way of calculating its value!

Convention: In what follows I shall use *cube*, *volume* and *solid parallelogram* as a sort of dimension-free shorthand for the corresponding n-dimensional concepts.

DETERMINANTS AND MATRIX MULTIPLICATION. From the transformation point of view, matrix multiplication corresponds to function composition. If we use this to calculate the magnification factor for AB, we get the product rule

$$\det(AB) = (\det A)(\det B).$$

Now suppose that A is an invertible $n \times n$ matrix. Since the transformation corresponding to the $n \times n$ identity matrix I clearly leaves everything where it is, we have $\det I = 1$. Hence the product rule shows that $(\det A^{-1})(\det A) = 1$. So $\det A$ must be nonzero and

$$\det(A^{-1}) = \frac{1}{\det A}.$$

What if A is not invertible? Since the reduced row echelon form of A is not the identity matrix, the system $A\mathbf{x} = \mathbf{0}$ has an infinite number of solutions, so there must be some nonzero $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$. Now imagine A transforming a cube having one of its sides parallel to $\mathbf{x}$. Since one of the edges of the image solid parallelogram has zero length, its volume must be zero. Hence $\det A = 0$. This means that the determinant can be used to test the invertibility of a matrix:

$$A \text{ is invertible if and only if } \det A \neq 0.$$

AN UNEXPECTED DEVELOPMENT. We want a practical way of evaluating $\det A$. We are going to use column operations (and later, row operations) to simplify the matrix. The idea here is that column operations on the matrix A correspond to geometric operations on the solid parallelogram that we have used to define $\det A$. Hence we can calculate the effect of these operations on the appropriate volumes.

We begin with the most innocuous-looking of the elementary column operations.

Swapping two columns: It is tempting to think that this operation does not change the image, and so the determinant is not affected. However, it turns out that we have to choose between this idea and the equally attractive idea that $\det A$ is an additive function if we let just one column vary at a time. For example, letting just the third column vary would give

$$\det(\mathbf{a}|\mathbf{b}|\mathbf{c} + \mathbf{d}) = \det(\mathbf{a}|\mathbf{b}|\mathbf{c}) + \det(\mathbf{a}|\mathbf{b}|\mathbf{d}).$$

As Figure 2 shows, this too is geometrically desirable. For a two dimensional picture of the same rule see [**O**, page 9]. Sometimes a three dimensional picture helps bring out the essentially two dimensional nature of the situation!

What is the connection between these ideas? Notice first that if the matrix A has two columns the same, then $\det A = 0$ since the matrix is not invertible or, more geometrically, since the image of the unit cube collapses because two of its edges coincide. Now suppose that we want to swap the first two columns of the matrix $A = (\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3)$. Then because of the additivity

$$\begin{aligned}
0 &= \det(\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_3)\\
&= \det(\mathbf{a}_1|\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_1 + \mathbf{a}_2|\mathbf{a}_3)\\
&= \det(\mathbf{a}_1|\mathbf{a}_1|\mathbf{a}_3) + \det(\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_1|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_2|\mathbf{a}_3)\\
&= \det(\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3) + \det(\mathbf{a}_2|\mathbf{a}_1|\mathbf{a}_3).
\end{aligned}$$

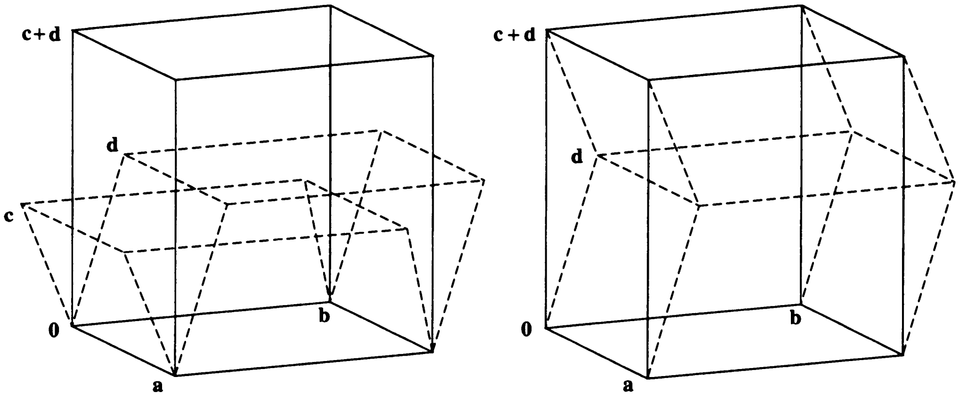

Figure 2. (*Left*) Images of the unit cube under the matrices $(\mathbf{a}|\mathbf{b}|\mathbf{c})$, $(\mathbf{a}|\mathbf{b}|\mathbf{d})$ and $(\mathbf{a}|\mathbf{b}|\mathbf{c}+\mathbf{d})$. (*Right*) Translating the image corresponding to the matrix $(\mathbf{a}|\mathbf{b}|\mathbf{c})$ shows that the combined volumes corresponding to the matrices $(\mathbf{a}|\mathbf{b}|\mathbf{c})$ and $(\mathbf{a}|\mathbf{b}|\mathbf{d})$ equal the volume corresponding to the matrix $(\mathbf{a}|\mathbf{b}|\mathbf{c}+\mathbf{d})$.

Hence

$$\det(\mathbf{a}_1|\mathbf{a}_2|\mathbf{a}_3) = -\det(\mathbf{a}_2|\mathbf{a}_1|\mathbf{a}_3),$$

and so swapping two columns of the matrix changes the sign of the determinant. This means we need to consider the possibility of some volumes being negative when we calculate the magnification factors for $\det A$. This unexpected development forces us to ask: were the pictures misleading?

Not really! It is more a question of us not noticing a subtle feature of the diagrams. Something similar happens when you first meet the integral as a way of calculating areas. There again you begin by imagining all areas are positive. But you have to allow for negative areas when you agree to rules like

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

Furthermore, this rule is forced on us (as it was with determinants), if we want an additivity formula

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx$$

to hold, even in situations (like $a = c$) that probably were not envisaged when you first tried to calculate areas.

What subtle feature have we missed so far? Clearly a positive determinant must corrąpond to a particular order of the columns of A. It makes sense that the identity matrix I ought have $\det I = +1$, so we should be able to observe the geometric effect of other column orders by swapping columns in I. Thus, in the 2×2 case, we should compare the effects of I and the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, as in Figure 3.

Although the image still has the same area as the original, it is actually a reflection of the original. We say that the transformation has *reversed the orientation* of the original. In general, a negative determinant indicates that a reflection is part of the action of the matrix.

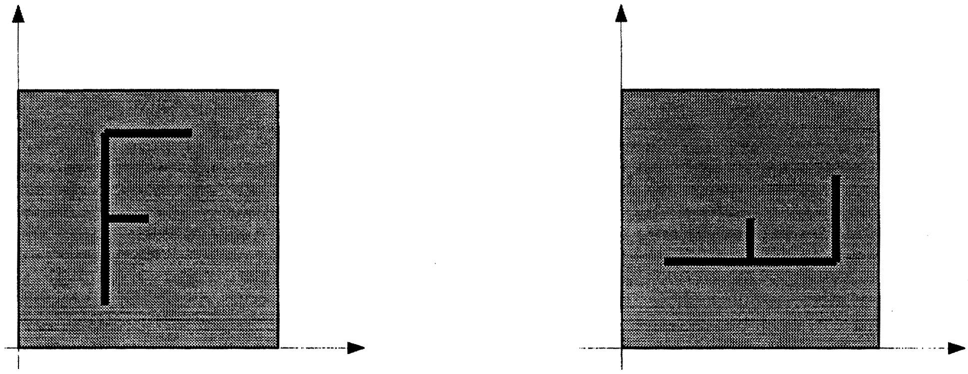

Figure 3. Images of the unit square under the action of the identity matrix (*left*) and the matrix $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ (*right*).

OTHER COLUMN OPERATIONS

Multiplying a column by a scalar: Since this operation scales one edge of the image, leaving all the other edges constant, the volume is scaled by the same amount. For example,

$$\det(\mathbf{a}_1 | s\mathbf{a}_2 | \mathbf{a}_3) = s \det(\mathbf{a}_1 | \mathbf{a}_2 | \mathbf{a}_3).$$

Notice that if the scalar is negative, then the orientation of the transformation is reversed.

Adding a multiple of one column to another: This operation produces a shear in the two dimensional picture corresponding to the two affected columns (see Figure 4), so there is no change in the volume or in the orientation. For example,

$$\det(\mathbf{a} | \mathbf{b} + s\mathbf{a} | \mathbf{c}) = \det(\mathbf{a} | \mathbf{b} | \mathbf{c}).$$

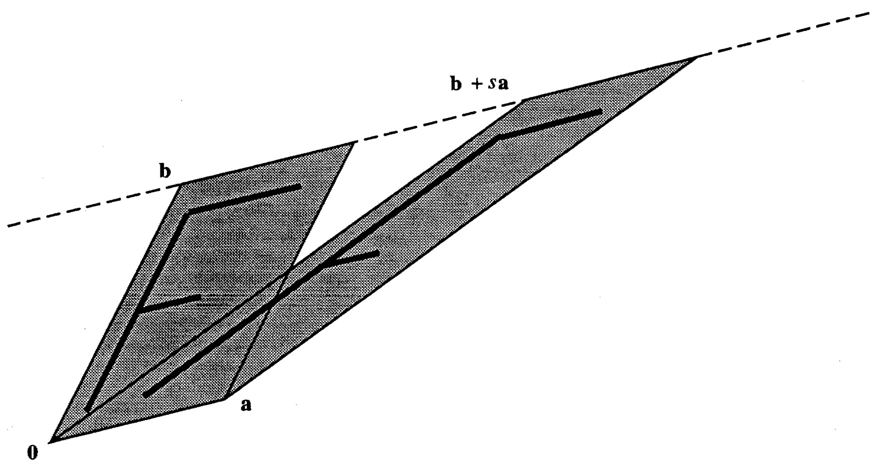

Figure 4. Images of the unit square under the matrices $(\mathbf{a}|\mathbf{b})$ and $(\mathbf{a}|\mathbf{b} + s\mathbf{a})$.

DETERMINANTS AND TRANSPOSES. A remarkable consequence of the product rule is that

$$\det A^T = \det A.$$

At this stage there does not seem to be a simple geometric reason for this result. Except for some simple matrices, there is no obvious geometric relationship between the columns of A and those of A^T. Consequently the explanation offered here will be essentially algebraic. However, as we shall see shortly, once the students have learned about eigenvalues and eigenvectors, we can give genuinely geometric reasons for this formula.

The simplest case is when A is not invertible. Then neither is A^T and both matrices have zero determinant.

So suppose that A is invertible. Then A is a product of elementary matrices. Because of the product rule, it follows that we just need to show that $\det E^T = \det E$ for any elementary matrix E. Now there are three different types of elementary matrix:

1. If E corresponds to the operation of multiplying the ith row by a nonzero scalar s, then E is just the identity matrix with its ith diagonal entry changed to s. So $E = E^T$ and clearly $\det E^T = \det E$.
2. If E corresponds to swapping two rows, then E is got by swapping the same rows of the identity matrix, and again $E = E^T$. So $\det E^T = \det E$ in this case too.
3. Suppose that E corresponds to the operation of adding s(row i) to row j. This affects only the ith and jth columns of E or E^T, so it is enough to look at what is happening to these two columns in the pictures for the transformations E and E^T. This makes it essentially a 2×2 matrix problem, comparing the determinants of the matrices

$$\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 \\ s & 1 \end{pmatrix}.$$

In this case, as Figure 5 shows, the corresponding parallelograms are simply reflections of one another. Furthermore, again as in Figure 5, the images still have positive orientation, and so once again $\det E^T = \det E$.

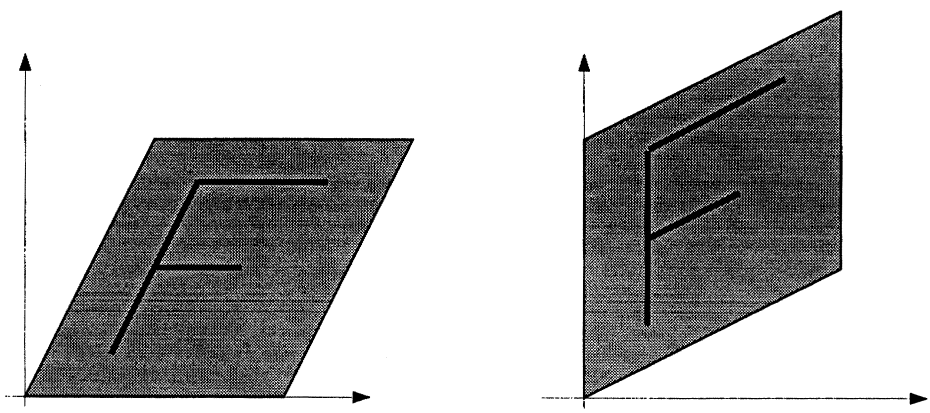

Figure 5. Images of the unit square under the transformations $\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$ (*left*) and $\begin{pmatrix} 1 & 0 \\ s & 1 \end{pmatrix}$ (*right*).

CALCULATING THE DETERMINANT. There is nothing new here. We have seen how column operations affect the determinant of a matrix, and we have also seen that $\det A^T = \det A$. Since column operations on A^T are exactly the same thing as row operations on A, this means that we also know the effect of row operations on the determinants. So we can use row operations (or column operations) to reduce the matrix to upper triangular form. Furthermore, we can use the same sort of operations to see that

If A is triangular, then $\det A$ is the product of the entries on the main diagonal of A.

Putting these two steps together gives the most efficient way of calculating $\det A$.

This completes my shopping list of "what every first year student should know about determinants," but of course I have left out several much-loved topics. Peer pressure ensures that I do at least mention them.

FORMULAS. What about a formula for $\det A$? We have now reached the stage where we have enough properties to find out what the formula must be in terms of the entries of A. These calculations also turn up some new ways of finding the determinant: the Laplace expansions of $\det A$ along the various rows and columns of A. See [**St**, pages 223–227] for one way of doing this.

The Laplace expansions of $\det A$ show that the adjugate (or classical adjoint) matrix, $\operatorname{adj} A$, constructed from the cofactors of A, satisfies

$$A \cdot (\operatorname{adj} A) = (\operatorname{adj} A) \cdot A = (\det A) I.$$

Hence as long as $\det A \neq 0$ we have the formula

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A.$$

This formula leads to Cramer's Rule for solving the system of linear equations $A\mathbf{x} = \mathbf{b}$. See [**St**, pages 231–233] for details.

Notice that none of these formulas is of much practical use, except perhaps in the 2×2 case. Their main value lies in their theoretical applications. In practice, $\det A, A^{-1}$, and the solution to $A\mathbf{x} = \mathbf{b}$ should all be found by using row operations (and/or column operations, if that makes sense).

EIGENVALUES, EIGENVECTORS AND DETERMINANTS. Eigenvalues and eigenvectors are another way of looking at the magnification properties of a linear transformation $\mathbf{x} \to A\mathbf{x}$. This time we look at linear magnification (the eigenvalues λ) along those directions (the eigenvectors $\mathbf{x}$) that are preserved by the transformation. Algebraically, these are related by $A\mathbf{x} = \lambda\mathbf{x}$.

In the case where the matrix A is diagonalizable (as is the case when the real matrix A is symmetric, for example), this gives another picture of the effect of A on volumes. Figure 6 shows a typical situation in the case where A is a 2×2 matrix.

So, at least in the diagonalizable case, we see that

$$\det A = \text{the product of the eigenvalues of } A.$$

Anecdote: Last time I taught this topic, I drew the picture in Figure 6 to show how the eigenvalues and eigenvectors could be used to describe the whole transformation. I did not mention determinants as I did not need the preceding displayed formula for later work in the course. However, one student came to me at the end of the session and asked whether the picture meant that $\det A$ had to be the product of the eigenvalues of A. The algebraic proof of this fact is simple,

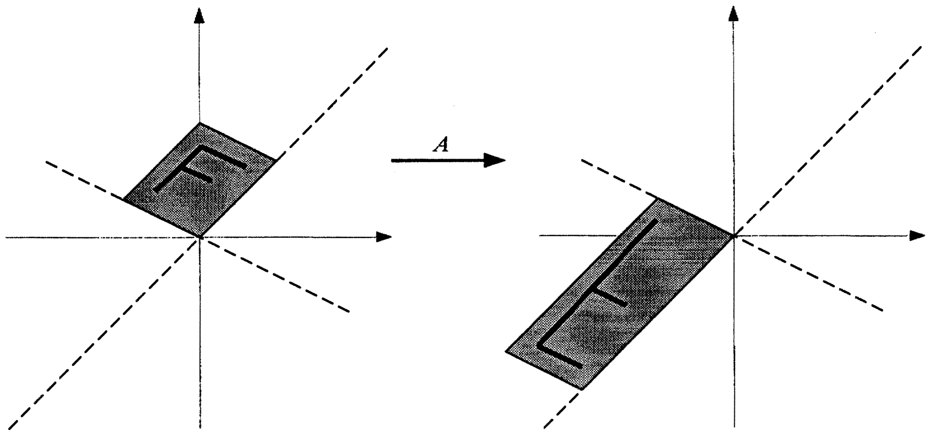

Figure 6. Effect of a matrix A with eigenvalues $\lambda = 1, -2$ and eigenvectors parallel to the dashed lines.

too (in the diagonalizable case), yet I've never had a student make the same connection when I have given an algebraic treatment of determinants.

DETERMINANTS AND VOLUMES OF ELLIPSOIDS. An n-dimensional ellipsoid is determined by an equation of the form $\mathbf{x}^T M \mathbf{x} = 1$ where M is a positive definite (real symmetric) matrix. The determinant can be used to find the (positive!) volume of this ellipsoid. Using a suitable rotation of axes we may assume that M is a diagonal matrix without altering det M. So the equation of the ellipsoid is

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \cdots + \lambda_n x_n^2 = 1$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of M. This is the same as

$$\frac{x_1^2}{\left(1/\sqrt{\lambda_1}\right)^2} + \frac{x_2^2}{\left(1/\sqrt{\lambda_2}\right)^2} + \cdots + \frac{x_n^2}{\left(1/\sqrt{\lambda_n}\right)^2} = 1,$$

which corresponds to a unit n-sphere that has been scaled by $1/\sqrt{\lambda_1}$ in the first coordinate direction, by $1/\sqrt{\lambda_2}$ in the second coordinate direction, and so on. Hence the volume of the ellipsoid must be

$$\left(\frac{1}{\sqrt{\lambda_1}}\right) \cdots \left(\frac{1}{\sqrt{\lambda_n}}\right) V = \frac{V}{\sqrt{\lambda_1 \cdots \lambda_n}} = \frac{V}{\sqrt{\det M}},$$

where V is the volume of the unit n-ball. For example, in the 2×2 case, where the unit disk has area π, the ellipse determined by $\mathbf{x}^T M \mathbf{x} = 1$ has area $\pi/\sqrt{\det M}$. Similarly, in the 3×3 case, the ellipsoid determined by $\mathbf{x}^T M \mathbf{x} = 1$ has volume $\frac{4}{3}\pi/\sqrt{\det M}$.

A GEOMETRIC VIEW OF THE DETERMINANT OF THE TRANSPOSE. We are now ready to see a more geometric explanation of the rule

$$\det A^T = \det A.$$

We can assume that A is an invertible matrix. The key idea is to evaluate the volume magnification of the transformation $\mathbf{x} \to A\mathbf{x}$ by looking at its effect on the unit n-sphere. It ends up being simpler to find the *inverse* image of the unit n-sphere $\mathbf{z}^T \mathbf{z} = 1$ (Figure 7).

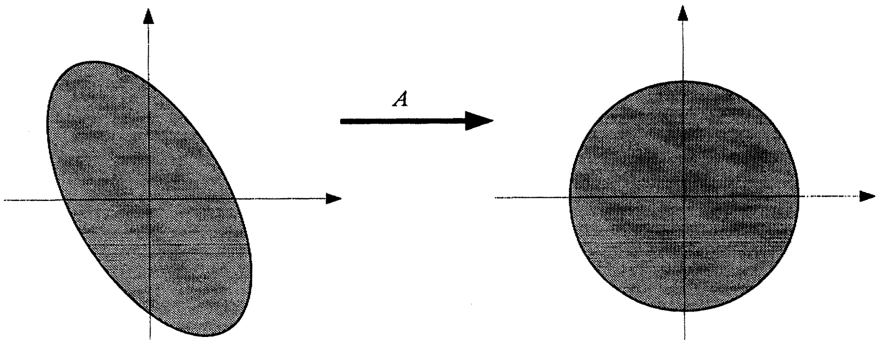

Figure 7. The inverse image under A of the unit n-sphere.

The points $\mathbf{x}$ mapping onto this sphere clearly come from the ellipsoid

$$1 = (A\mathbf{x})^T A\mathbf{x} = \mathbf{x}^T(A^TA)\mathbf{x}.$$

Now A^TA is a positive definite matrix, so the preceding discussion on volumes of ellipsoids can be applied here. Thus the inverse image has (positive) volume

$$\frac{V}{\sqrt{\det A^TA}}$$

where V is again the volume of the unit n-ball. Hence the (signed) magnification factor of the transformation $\mathbf{x} \to A\mathbf{x}$ is given by $\pm\sqrt{\det A^TA}$. By our definition of $\det A$ we thus have $\det A = \pm\sqrt{\det A^TA}$. Squaring and using the product rule to cancel a factor of $\det A$, we get the desired formula.

REFERENCES

[AK] Agnew, J. L., and Knapp, R. C., *Linear algebra with applications*, 3rd ed., Brooks/Cole, 1989.

[BK] Blum, W., and Kirsch, A., Preformal proving: examples and reflections, *Educ. Stud. Math.* **22** (1991) 183–203.

[DL] Damiano, D. B., and Little, J. B., *A course in linear algebra*, Harcourt, Brace and Jovanovich, 1988.

[FB] Fraleigh, J. B., and Beauregard, R. A., *Linear Algebra*, 2nd ed., Addison-Wesley, 1990.

[N] Nicholson, W. K., *Elementary linear algebra with applications*, 2nd ed., PWS-Kent, 1990.

[O] Open University, *Determinants and eigenvalues*, (Linear Mathematics, Unit 5), The Open University Press, 1972.

[Se] Sernesi, E., *Linear algebra: a geometric approach*, Chapman and Hall, 1993.

[St] Strang, G., *Linear Algebra and its Applications* (3rd edition), Harcourt Brace Jovanovich, 1988.

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch
New Zealand
j.hannah@math.canterbury.ac.nz