Matrix Algebra Review
Instructor(s): Dr. Robert LaBudde
Last updated from July 2015 course

Note: These discussions have been selected from prior courses and lightly edited to correct typos and remove student names. A useful way to use this resource is by browsing, or via the index.

WEEK 1

Subject: Inner product
Participant's Question:
Your text cites "Because of the way in which the inner product is defined as a scalar, the inner product can actually be computed between any of 2 row vectors, 2 column vectors, a row and a column vector and a column and a row vector, so long as
both vectors have the same order (number of elements)."
Must I understand this only applies if one of the two is transposed, or are you referring to the way it's written our in R?
Instructor's Response:
To be punctilious, the inner product should be an operator in a particular vector space, which would be either row vectors or column vectors, but not both. That's the way it's taught in math. However, the practical definition of an inner product by its formula is the same as is used for matrix multiplication, and it's convenient to treat this operation as generalized this way. In a similar manner, most applications treat a 1x1 matrix as a scalar, although from a math point of view scalars and matrices are different objects.

We're teaching matrix algebra here the way it's used in statistical applications, not as abstract math based on deductive logic from axioms and definitions. This saves a lot of complicated and confusing justifications and works out okay at the end. I do, however, point out these conceptual breaches as they occur.

Finally, R sometimes maintains a distinction between row and column vectors and sometimes not. This can be annoying when trying to get results from a vector/matrix operator. Also, R treats matrices and vectors as different objects.

If **a** is a row vector and **b** is a column vector, then the matrix product **a b** has the same numerical value as **a' . b** or **a . b'**, except that the former is really a 1x1 matrix, not a scalar. In this sense (the formula for value), the three operations could be considered equivalent, although they result in different object classes strictly speaking.

In R there is only matrix multiplication '%*%':

```
> a<- t(c(1,2))
> b<- c(3,4)
> str(a)
num [1, 1:2] 1 2
> str(b)
```

```
num [1:2] 3 4
> a%*%b
[,1]
[1,] 11
> str(a%*%b)
num [1, 1] 11
> d<- c(5,6)
> b%*%d
[,1]
[1,] 39
> str(b%*%d)
num [1, 1] 39
```

Note that "a'" is treated as a matrix (transpose forces conversion), and all results using '%*%' also result in matrix values.

Subject: 'uv' vs. 'vu' and knowing which is implied as an outer product
Participant's Question:
I'm looking at Exercises 1 & 2 in the class notes.
1. "Compute u v"
2. "Compute v u"
The hint says the first one is a "dot-product" or inner product the second an outer product. What is supposed to be my cue here, since the notation is the same? Am I supposed to infer the second instance is v X u because v is named first and v is a column vector?
Instructor's Response:
We could have used explicit terminology to make clear the ambiguity: '**u** . **v**' for the scalar (inner) product, and '**u** x **v**' for the matrix (outer) product. (The syntax '**u** . **v**' is why the first is called the 'dot' product.)
However, only one type of product makes sense in each case. If **u** is a row vector and **v** is a column vector, the **u v** must be the inner product. Similarly **v u** must be the outer product.
Even simpler, it doesn't make any difference in any case how you interpret the formula, so long as you compute elements using the standard matrix multiplication of rows and columns. This automatically gives the correct answer.
The only 'gotcha' here is that, strictly speaking, the inner product gives a scalar result, but the matrix multiplication formula gives a matrix result. You need to keep separate the scalar 2 (which is of class numeric) and [2] which is a 1x1 matrix. Conventionally 1x1 matrices are treated as scalars, even though properly speaking they are matrices. The application use indicates which is the concept needed. You have to watch out for this problem in using software, which is unforgiving on typing errors.
Participant Continued…
So whenever we have a row X column situation, the convention is to go ahead and sum numbers in the resulting single column into a scalar.
Instructor's Response:
Yes, that's correct. Although you should start paying attention to the details of notation to avoid difficulties later. Either say 'u v' or 'u. v' for the inner product. Saying 'u X v' would suggest an outer, rather than inner, product; the way notation is used in matrix algebra.
You should also note that, in R, u*v actually computes a vector whose elements are the

product element-wise of the two vectors:

```
> u<- c(1,2,3)
> v<- c(4,5,6)
> u*v
[1] 4 10 18
> u%*%v
[,1]
[1,] 32
> str(u%*%v)
num [1, 1] 32
> class(u%*%v)
[1] "matrix"
```

Notice the 'gotcha' from R: u%*%v is a 1x1 matrix, not a scalar, so R is doing a matrix product, not an inner product. Sometimes this causes problems when the result is used in function calls in R. You can force the result to a scalar by using the as.numeric() function on the expression:

```
> as.numeric(u%*%v)
[1] 32
> class(as.numeric(u%*%v))
[1] "numeric"
> str(as.numeric(u%*%v))
num 32
```

Luckily most of the time R coerces types to fit the situation, so you don't have to worry about this kind of thing all the time. But it does appear to crop more with vectors, matrices, and data frames than any other types in R.

Subject: Trace
Participant's Question:
What is the significance of the trace matrix and its role in multivariate statistics and econometrics?
Instructor's Response:
The trace of a matrix is the sum of its diagonal elements, and is primarily useful for square matrices. It is useful as a scalar measure of a property of a matrix, so it's easy to deal with, being just a number. There are certain operations on matrices that leave the trace unchanged, making it a property invariant to these operations. (The trace of a linear combination of matrices is the linear combination of the traces, e.g.)
To be more specific, consider the multivariate covariance matrix **V**, whose diagonal elements are the variances of the variables, and the off-diagonal elements are the covariances between them. tr(**V**) then is the sum of the variances of all the variables. This obviously has something to do with how much the variables vary, a smaller value meaning less total variance and a higher more variance in total.
Much of multivariate statistics is, besides modeling, trying to find scalar statistics that measure the effect size of a matrix problem. There are only a few scalar values (trace, determinant, eigenvalues) that measure a matrix overall, and these are frequently found as the basis of statistical tests. One of these, e.g., is 'Pillai's trace', which is used to test the

significance of multivariate analysis of variance ('MANOVA'). The 'Lawley-Hotelling trace' is also used for this purpose.

Subject: Singular
Participant's Question:
What is meant by the term 'singular'? Is its meaning contextual? Is there just one definition?
Instructor's Response:
"Singular" is a term used in mathematics to indicate there is something special going on, typically bad.
In the case of matrices, a matrix is "singular" if it has no inverse and "nonsingular" if it does. This is a generalization of dividing by zero, which is a "singularity" in mathematical terms.

WEEK 2

Subject: Homogeneous system of equations
Participant's Question:
I don't understand at all the concept here of Homogenous System of Equations. I can do the steps and get an answer. But what the heck is it supposed to be telling me? Read the notes several times and the book on this concept, and it simply isn't computing. I don't get what the purpose is.
Instructor's Response:
A homogeneous system of equations is in the form $\mathbf{A}\,\mathbf{x} = \mathbf{0}$. Obviously $\mathbf{x} = \mathbf{0}$ is a solution. Normally the useful solutions are those $\mathbf{x}$ that make the result come out $\mathbf{0}$ without being all zero itself. If there are any such 'non-trivial' solutions, there will be an infinite number of them.

If $\mathbf{A}\,\mathbf{x} = \mathbf{0}$, this means the components of $\mathbf{x}$ in the solution set are not independent, so at least one of them can be eliminated by substituting a linear expression in terms of the others. I.e., you have too many unknowns with not enough equations to pin them down. So you have a 'degree of freedom' or more in choosing the components of $\mathbf{x}$. Whenever you have this kind of redundancy (or 'degeneracy' as it's called), you have to deal with it explicitly in subsequent analysis. In statistics you'll see either one of the variables eliminated before continuing, or you will see one or more Lagrange multipliers used if it's an optimization problem (like maximum likelihood). (If you don't know what a Lagrange multiplier is, it is explained in advanced calculus.)

In matrix algebra, if $\mathbf{A}\,\mathbf{x} = \mathbf{0}$ for any x except $\mathbf{x} = \mathbf{0}$, the rows of A are 'linearly dependent' on each other.
Participant Continued…
I'm not quite clear on the last comment.
"In matrix algebra, if A x = 0 for any x except x = 0, the rows of A are "linearly" dependent on each other."
So let's say we have some non-square matrix. And let's say we're trying to find the generalized inverse and you "think" you've already eliminated any linear dependence, are

you saying then that when solving for A x = 0, we should at that point only get a solution for x = 0? In other words, hypothetically, let's say I get another solution (say x = 2... doesn't matter... anything other than x = 0) that satisfies the equation A x = 0, then I still have linear dependence and need to go back and recheck my work?
Actually, I guess it doesn't matter if it's a square matrix or not. If I'm trying to find the generalized inverse (which in the case of a nonsingular square matrix will have a particular solution), if it turns out that A x = 0 has another other solution than x = 0 to satisfy the equation, then basically, I've made some mistake. Is that an accurate statement? Or am I still misunderstanding?

Instructor Continued…

1. Linear dependence and independence are discussed in section VIII of the notes.

2. If **A x** = **0** has any solution other than **x** = **0**, then A is 'singular' and it does not have an inverse. It cannot have an inverse if it is not square either. The best you can do then is a 'generalized inverse', which is available for all matrices whether square or not or singular or nonsingular.

Subject: Determinant

Participant's Question:

On determinants of a 3x3 matrix, you have the following statement (pg 41):

"Note that each term in eq.(XIII.6) is of the form $a1_{j1}\ a2_{j2}\ a3_{j3}$, where (j1 j2 j3) has the same elements as (1 2 3), but with a different order. The sign (+1 or –1) in front of each term in eq.(XIII.6) is +1 if {j1, j2, j3} is an even permutation of (1 2 3) and –1 if it is an odd permutation."

Honestly, it makes no sense. I simply don't know what that means. It's not keeping me from doing the homework, but I wouldn't be able to generalize, say a 4x4 matrix, based off of the above explanation. And, unfortunately, I'm not having any luck deciphering the generalized formula on the next page (XIII.7). Not because there's anything "wrong" with it, I just don't understand it. I'm not following the notation or the explanation. The book's explanation gets into a concept of a "cofactor" matrix, where I can determine the signs of the cofactors, but then it seems to use a different process (that I'm sure gets to the same result) to find the determinant. And it's discussion of the "weighted sum" in this context is totally lost on me.

Instructor's Response:

1. When you work out the determinant, you find that it is a sum of products of the elements of the matrix **A**, each product having exactly one term from each row and each column. But the subscripts are scrambled ('permuted'). Because of this, it's easy to write down a closed form expression for the determinant:

$\det(\mathbf{A}) = +/-a_{11}\ a_{22} +/- a_{12}\ a_{21}$

The only tricky question is what sign to use on each term. The answer is the 'sign' of the permutation for each product's 2nd subscript. The 'sign' of the permutation is how many transpositions of pairs of numbers are needed to get the numbers back into order, as a

power of -1. If this results in +1, the permutation is called 'even' (i.e., even number of transpositions needed), if -1, then 'odd'.

For {1,2}->{1,2}, no transpositions are needed, so the sgn{1,2} = $(-1)^0$ = +1.
For {2,1}->{1,2}, only 1 transposition is needed, so the sgn{2,1} = $(-1)^1$ = -1.

For a 3x3 matrix **A**:

det(**A**) = +/- a11 a22 a33 +/- a11 a23 a32 +/- a12 a21 a33 +/- a12 a23 a31
+/- a13 a21 a32 +/- a13 a22 a31

For example, sgn{1,3,2) = -1 (single transposition needed to get to {1,2,3}). Or sgn{2,1,3} = +1 (two transpositions needed to get to {1,2,3}).

This really is only useful theoretically or for very small order matrices. For larger matrices, row-reducing to triangular form and then multiplying the diagonal elements is more efficient computationally.

2. Like most things in mathematics, there are many different formulas or algorithms that can be used to get the same answers. Much of mathematics is spent proving that the different formulas always give equal answers. Looking at things in different ways may be confusing, but it leads to different generalizations, which is a 'good thing'.

3. PS. 'permutations' and 'combinations' and other such enumerative methodologies comprise a field of study called 'combinatorics'. This is included in statistics programs as part of probability theory. It is a difficult topic to grasp just because it is so simple and therefore abstract. It's so simple that it's actually an advanced topic!

Participant Continued…

I'm still missing something. I don't understand what you mean by "transpositions" of {1,3,2} or {2,1,3}. First, where are the numbers coming from? Are they the subscripts? And are they the row or column subscript? And what do you mean by "transposition". I know what the word means, and that's why this doesn't make sense. {1,3,2} requires "transposing" the 3 and the 2 to get to (1,2,3). But (2,1,3) requires transposing the 1 and the 2 to get to (1,2,3). Those both look like one transposition. But you say the first has 1 transposition, but the second has 2 transpositions. That simply does not compute to me. So I'm obviously not understanding what you mean by "transposition" in this context.

Instructor Continued…

1. In the formulas of eq.(XIII.6), it is the column subscripts which are being permuted. The ordered set {1,3,2} would refer to the column subscripts of a11*a23*a32.
2. A 'transposition' of {1,3,2} would be a single interchange of two of the values. For example {1,3,2}->{1,2,3} by 'flipping' or 'transposing' the last two values. You are right: It only takes one transposition to {2,1,3}->{1,2,3}. This was a mistake on my part. The example {2,3,1} takes 2 transpositions and has sgn = +1.
http://en.wikipedia.org/wiki/Transposition_%28mathematics%29#Transpositions

Subject: Echelon insight
Participant's Question:
Can you provide some insight or hints on how to analyze or view a matrix when determining what elementary function will create an echelon form of the matrix in question?
Instructor's Response:
Notes give a short algorithm to find the inverse of a matrix efficiently. You systematically reduce the original matrix to an identity matrix, and at the same time you apply the same operations to a starting identity matrix. When A becomes an identity matrix, the original identity matrix becomes the inverse of A. This is easiest done by hand by appending the identity matrix as extra columns of A, then carrying out all row operations on the 'augmented' matrix. At the end, these last columns contain the inverse. On a computer, it is possible to develop the inverse in place by careful bookkeeping, and at the end the inverse is present in the original matrix storage locations.

The steps are:
1. Start at the next row (first time, row 1). This is called the 'pivot' row.
2. Find the largest absolute value element in the column containing the diagonal element of the row (first time, column 1), only considering values at or below the diagonal.
3. Swap the two rows to get the largest value on the diagonal of the pivot row.
4. Divide the pivot row through by its diagonal element. The makes the diagonal element 1.
5. Make the pivot column all zeroes by subtracting multiples of the pivot row from the rows below the pivot row (all rows above already have zero in this column).
6. Repeat for the next column.
If there is an inverse, this process will find it.

Subject: Row Echelon form
Participant's Question:
Is there a built-in function in R for performing the Row Echelon Form?
I know it's not covered in your notes. But, what is the difference between Row Echelon Form and Reduced Row Echelon Form? Which one should we use?
Instructor's Response:
I did find one function rref() in package 'pracma':

http://127.0.0.1:14956/library/pracma/html/rref.html

```
> require('pracma')
> A<- matrix(runif(12), ncol=4) #random 3x4 matrix
> A
[,1] [,2] [,3] [,4]
[1,] 0.2949330 0.7573227 0.5697436 0.30476537
[2,] 0.2928260 0.0407312 0.7656623 0.02403664
[3,] 0.9526053 0.6764003 0.6333913 0.63614271
> rref(A) # reduced row echelon form
[,1] [,2] [,3] [,4]
[1,] 1 0 0 0.5646390
[2,] 0 1 0 0.3347692
[3,] 0 0 1 -0.2023606
```

'Row echelon form' has zeroes in rows and columns under the diagonal. 'Reduced row echelon form' has zeros in rows above the diagonal as well. For our purpose 'row echelon form' suffices to find the rank of a matrix, and requires fewer math operations to arrive at. Row echelon form can always be reduced further to reduced row echelon form.

Subject: Echelon
Participant's Question:
Can you provide a less technical definition for echelon form or a rewording perhaps?
Instructor's Response:
Less technical: A is in row echelon form if the principal diagonal has 1's (and then possibly 0's) on it, and 0's below it. A is in reduced row echelon form if the principal diagonal has 1's (and then possibly 0's) on it and 0's elsewhere in the columns of the principal diagonal.
Some people don't require that the numbers on the diagonal be 1 or 0. They allow non-zero values other than 1. (But these can always be divided through the row to make it 1, as we do here.)
See http://en.wikipedia.org/wiki/Row_echelon_form.

Subject: R question
Participant's Question:
Is there a built-in function in R that automatically creates an identity matrix as long as the user(programmer) gives the order of the matrix?
Instructor's Response:

```
diag(3) gives a 3x3 identity matrix.
> diag(3)
[,1] [,2] [,3]
[1,] 1 0 0
[2,] 0 1 0
[3,] 0 0 1
> diag(c(1,2,3))
[,1] [,2] [,3]
[1,] 1 0 0
[2,] 0 2 0
[3,] 0 0 3
> A<- matrix(1:9,ncol=3)
> A
[,1] [,2] [,3]
[1,] 1 4 7
[2,] 2 5 8
[3,] 3 6 9
> diag(A)
[1] 1 5 9
```

Subject: Application of Generalized inverse matrix
Participant's Question:
How are generalized inverse matrixes used in other fields and statistical applications? I remember using inverses of square matrixes and eigenvalues in my engineering classes but nothing about generalized inverses.
Instructor's Response:

Obviously generalized inverses are only useful when rank(A) is less than full.
A number of multivariate statistical methods use matrices that can be only expected to have partial rank, but you still want answers for the dimensions that are determined. This occurs most often in regression modeling.
As far as engineering and science are concerned, I have less knowledge. But here's an arcane example:
http://drs.nio.org/drs/handle/2264/954

Subject: Right Generalized Inverse
Participant's Question:
What do you mean by right generalized inverse? The book does not mention "right" so this confuses me.
Instructor's Response:
A generalized inverse has a different shape depending upon whether you want to multiple on the left (G A) or the right (A G). All of the examples in this course are on the right, so my use of it in the assignment is punctilious and redundant, so I see why it might be confusing. For the problem given, A is a 2 x 3 matrix, so the (right) generalized inverse G will have to be 3 x 2 to conform. If it were a (left) generalized inverse you wanted, it would have to be 2 x 3 instead.

Subject: Independent Matrices
Participant's Question: Can you clarify what is meant by:"In two dimensions (n = 2), if u1 and u2 are linearly independent, then they lie on a plane, but not on a line, so they are suitable for functioning as coordinate axes on the plane via eq.(VIII.2). In such a case, u1 and u2 are said to "span" the plane. If u1 and u2 are linearly dependent, however, they are collinear, and can't supply two independent directions to serve as coordinate axes."?

Instructor's Response: A 'vector' in analytical geometry is not a 'point', it is a directed line segment from one point (usually the origin) to another point. If the two line segments line on the same line, then they are 'collinear' (whence the term). If not, they are composed of two independent directions, and can therefore form the basis of 2-D coordinates. Just as two 'points' will always determine a line, two 'vectors', if independent, will always determine a plane.

'Vectors' as geometrical objects are extensively developed in physics and are used heavily in the theory of electric and magnetism, and in mechanics.

From a geometric point of view, two vectors are independent if they don't point in exactly the same direction.Orthogonality means the two vectors have no common direction (at right angle). $\mathbf{u} \cdot \mathbf{v} = 0$.If two vectors are 'dependent', or 'collinear', they lie on the same line, so one is a multiple of the other. $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}|\,|\mathbf{v}|$.If two vectors are not 'dependent', they are 'independent'.If u and v are two 'independent' vectors, then let w be defined by

$$\mathbf{w} = \mathbf{u} - (\mathbf{u} \cdot \mathbf{v})\,\mathbf{v}$$

Then **w** is the part of **u** that is not dependent on **v**, and **w . v** = 0 (orthogonal).Generally we can transform an independent set of vectors this way into a set of orthogonal vectors. It's called 'Gram-Schmidt orthogonalization'.

WEEK 3

Subject: Method to find Eigenvectors
Participant's Question:
Which is the preferred method for finding eigenvectors? The method on page 4 or page 6 (which uses the generalized inverse of the original matrix)? The later seems more time consuming since we have to calculate the generalized inverse separately.
Instructor's Response:
For hand calculation, there is no preferred method. The generalized inverse method always works, but is a lot of trouble for just a 2 or 3 dimensional matrix.
The simplest method is just to write out the **A x** = lambda **x** equations and solve then by hand. Once you find an **x** that works, divide by the norm of x to make a unit vector.
The first step is obviously to find the values of lambda (the eigenvalues).

Subject: norm of a vector
Participant's Question:
On page 5 what does |u| mean (on the third line of text)? Actually it's on the second line of text on page 5.
Instructor's Response:
|**u**| or ||**u**|| is the 'norm' of a vector, or the square root of the inner product **u** . **u**.
If **u** = [x y]', then | **u** | = sqrt(x*x + y*y).

Subject: Unit eigenvectors
Participant's Question:
I wonder if a simple formula is available for how to express eigenvectors as 'unit eigenvectors' and 'normalized unit eigenvectors'.
Instructor's Response:
To 'normalize' a vector into a 'unit vector', all you need to do is:
1. Add the squares of the elements of the vector.
2. Take the square root of this number.
3. Divide each element of the vector by this result.
When you are done the sum of the squares of each element add up to unity.

Subject: Matrix Algebra
Participant's Question:
Just from reading the class text and notes it's clear how astonishingly and richly nuanced are the ideas of order and systematic relations that underlie and animate matrix algebra. It's really quite an amazing view, even from just lifting the hood as I'm doing here in this class.

Could you guys ramble on a bit about the evolution of matrix algebra. Nothing long winded, but just a paragraph or so. Did it start with Gauss? The way calculus started with Newton and Liebnitz. Were there any other mathematicians who advanced the subject and contributed new ideas to it? Is it an ongoing body of thought or is it pretty much a completed work at this point? Any thoughts would be appreciated and if it's a long story, don't sweat it too much.

Instructor's Response:
Matrix algebra is a subtopic of linear algebra, which has flourished in development in the 19th and 20th centuries.
An interesting sidenote in the history is that the idea of directed 'vectors' and the associated 'vector algebra', including inner, outer and cross products, was invented by a famous American physicist, J. Willard Gibbs in the 19th century. This algebra became the basis for electromagnetic theory (Maxwell's equations are given as vector-valued equations) and thermodynamics (credited in its invention to Gibbs).
Linear algebra is now the underlying basis of all physical disciplines (physics, chemistry, engineering) as well as a large part of modern mathematics and statistics. It is the language of science.
See, e.g.:
http://en.wikipedia.org/wiki/Linear_algebra
http://link.springer.com/chapter/10.1007/978-0-8176-4685-1_5#page-1
http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CC0QFjAB&url=http%3A%2F%2Fwww.math.utah.edu%2F~gustafso%2Fs2012%2F2270%2Fweb-projects%2Fchristensen-HistoryLinearAlgebra.pdf&ei=WgeCVI6KNMqpNv-ShAg&usg=AFQjCNHSjZEXti6ewFd08aRvkWaRK9VQmQ&sig2=j7tod6kbr-wL8h2R2-9ADg&bvm=bv.80642063,d.eXY

Subject: Null Space
Participant's Question:
Can you please explain what is a Null Space? And its significance in system of equations and matrices in general?
Instructor's Response:
The 'null space' or 'nullity' of a matrix is the set of all vectors **x** which satisfy **A x** = **0**.
You will note that an easy way to get this is to find all eigenvalues and eigenvectors of **A**, and then the eigenvectors corresponding to an eigenvalue of 0 span the null space of **A**.
Any linear combination of these eigenvectors is in the nullity of **A**.
Because **A x** = **0**, any expression that is multiplied by **A** will have the same value if an arbitrary **x** from the nullity is added to it.
Subject: Applications
Participant's Question:
While the introduction and navigation of concepts is somewhat adequately dealt with in the book, I am still failing to see the connection to real-world applications. Could you perhaps illustrate a few examples (or provide references to material) of the applications for determinants, eigenvalues and eigenvectors and principal components in statistics...
Instructor's Response:

1. DETERMINANTS: The multivariate normal distribution involves the determinant of the covariance matrix (generalized variance) in its definition (see http://en.wikipedia.org/wiki/Multivariate_normal_distribution). Consequently the determinant (and the trace) appear throughout multivariate statistics, particularly in statistical tests related to the structure of the covariance or correlation matrices. This includes tests of 'sphericity' (equal variances and uncorrelated variables), tests of 'compound symmetry' (equal correlations) and tests of a particular form of the covariance matrix (see, e.g., http://www.graphpad.com/faq/viewfaq.cfm?faq=1500).
2. EIGENVALUES/EIGENVECTORS & PCA: Principal components are used in statistics generally for two purposes: 1) to reduce dimensionality before plotting or carrying out multivariate analysis; or 2) in principal components form of exploratory factor analysis. Using the first two or first three principal components allows visualization in 2D or 3D plots of what would be otherwise high dimensional data with no easy way to graphically display. Also, using the first few principal components as a transformed dataset frequently makes such methods as cluster analysis a lot more effective. Exploratory factor analysis is used in social sciences to investigate models that give rise to observed correlation structures and to identify 'latent' or unobserved variables (e.g., 'intelligence') that create the observed structure.
'Statistics' is a collection of methodologies that condense a large pile of data into a much smaller handful of 'statistics' that contain all or almost all the same information as the large pile did. We use such 'statistics' to test various models that make our data even simpler in conceptual underpinnings. In the same way PCA reduces a large number of variables into a handful that contain almost the same information. Otherwise multivariate problems can be too complex to model in a stable way or understand properly. Multivariate problems are said to suffer from the 'curse of dimensionality'.
Eigensets are also used in canonical correlation analysis and canonical variates analysis in multivariate analysis of variance ('MANOVA').
When you take a course in multivariate statistics or data mining, you will encounter determinants and eigenvalue analysis frequently as methods and algorithms are explained.
One obvious application that is very practical is that of 'credit score', which typically could be the first principal component score based on all the dozens of credit-related variables measured. Another is Olympics scoring of decathlons, which works out to be statistically equivalent to the first principal component score of the 10 event scores.

Subject: Conceptual questions
Participant's Question:
1) Why use the 'unit norm' and not some other combination of eigen values? Is that because scaling (at the level of principal components) becomes easier down the line?
2) What happens when there is a multiplicity of eigen values, i.e in terms of principal component analysis?
3) Somewhat of an un-related question is:
Is there a reason to choose one type of matrix over the other in the text book example (page 89), i.e choosing a sum of squares and products matrix over a variance-co-variance matrix?
Instructor's Response:

1. The virtue of normalizing the eigenvectors by using the L(2) norm (i.e., defining the norm as the root of the inner product with itself) is that the eigenvector matrix is 'orthogonal' or 'orthonormal', which makes its transpose its inverse for real-valued symmetric matrices A. This is very useful, and also means that the eigenvector matrix provides a length-preserving rotation of coordinate axes when given a geometric interpretations.
2. Multiple eigenvalues are a complexity that we don't deal with in this course. They are not that hard to deal with, but present additional ambiguity. E.g., an eigenvalue of multiplicity two (two equal real roots of the characteristic equation) is associated with two different eigenvectors, which can be chosen to be orthogonal to each other and normalized. These eigenvectors are ambiguous up to an orthogonal transformation. In PCA, this creates uninterpretability. In factor analysis, the rotation is chosen to simplify interpretation.
3. Typically the correlation matrix is used for PCA (to put all variables on equal scales), sometimes the covariance matrix is used (if all variables are already on near equal scales). It is a rarity to use the moment matrix (sum of crossproducts matrix), as the text indicates. Because the Table 4.1 example has zero means for the two variables, the moment matrix W = (n-1) S, where S = covariance matrix and (n-1) = d.f. Using R:

```
> #MAR: p. 89
> A<- matrix(c(0.9,0.2, 0.8,0.4, 0.5,0.4, 0.2,0.6, 0.2,0.7, 0.0,0.1,
0.0,-0.1,
+ -0.2,-0.7, -0.2,-0.6, -0.5,-0.4, -0.8,-0.4, -0.9,-0.2), ncol=2,
byrow=TRUE)
> A
[,1] [,2]
[1,] 0.9 0.2
[2,] 0.8 0.4
[3,] 0.5 0.4
[4,] 0.2 0.6
[5,] 0.2 0.7
[6,] 0.0 0.1
[7,] 0.0 -0.1
[8,] -0.2 -0.7
[9,] -0.2 -0.6
[10,] -0.5 -0.4
[11,] -0.8 -0.4
[12,] -0.9 -0.2
> colMeans(A)
[1] 0 0
> W<- t(A)%*%A
> W
[,1] [,2]
[1,] 3.56 1.92
[2,] 1.92 2.44
> S<- cov(A)
> S
[,1] [,2]
[1,] 0.3236364 0.1745455
[2,] 0.1745455 0.2218182
> 11*S
[,1] [,2]
[1,] 3.56 1.92
```

```
[2,] 1.92 2.44
```

I believe the book uses W instead of S to avoid small numbers (the factor of 11 keeps the eigenvalues near 1.0). The two variables are on nearly the same scale (variances of 0.32 vs 0.22), so PCA on the S would be logical. The PCA on W is the same as that on S, except that the eigenvalues are 11 times larger.

Subject: Lesson 3 Question
Participant's Question:
In the notes (and likewise using R) the presented solution for $[1, 1][x, y]^t=0$ is $[-1, 1]^t$. It seems to me that $[1, -1]^t$ is equally valid.
Is it the convention to have the positive on the bottom?
Instructor's Response:
Great question! Eigenvectors are ambiguous up to a factor of +/-1 as defined. Some people add the convention that the first element be positive, but not all people do. R appears to output whatever sign the calculations lead to without using any convention, so it's 'random'.

Subject: Diagonalization
Participant's Question:
The book spend a great deal of time, pages 82-88, discussing diagonalizing matrices. I find this section confusing and do not understand the motivation behind it. Can you explain further?
Instructor's Response:
'Diagonalization' means that, under some regularity conditions on multiple eigenvalues, A can be shown 'similar' to a diagonal matrix D which has the eigenvalues on the diagonal. That is,
A U = U D => U^(-1) A U = U^(-1) U D = D
So multiplying A on the left by U^-1 and the right by U (where U is the matrix whose columns are the eigenvectors) turns A into the diagonal matrix of eigenvalues D. (The regularity conditions ensure U is invertible.) It is also true that
A = U D U^-1
This is not of great use for general square matrices A. However, when A is a symmetric matrix, U is 'orthogonal', which means its inverse exists and is equal to its transpose: U^-1 = U'.
So if A is symmetric, we can say that, given the eigenvalues and the eigenvector matrix U, then U' A U = D. It also means that A = U D U', which is the 'spectral decomposition' of A. Writing out the details of U D U',
A = g1 u1 u1' + g2 u2 u2' + ... + gn un un'
where the g's are the eigenvalues and the u's are the eigenvectors.

Subject: Generalized Sample Variance
Participant's Question:
I have another question. It is in relation to lesson 2; the application of determinants in statistics. I am having difficulty understanding "generalized sample variance". What does

it mean to be "proportional to the aquare of the volume enclosed by the n vector deviations from the vector mean in the data". What does "degenerate" mean?

Instructor's Response:

The "generalized sample variance" is defined as |S| (the determinant of the sample covariance matrix).

The matrix X of the sample data can be considered as a collection of vectors starting at 0 and pointing to points on an enclosing surface (an ellipsoid, in fact). The covariance matrix describes the overall spread of that
shape in each direction.

The equation (x − xbar)' S^(−1) (x - xbar) = a^2 defines a locus of all points equidistant from xbar (weighted in each dimension by S^(-1)). This is an ellipsoid with semi-axes proportional to the square roots of the eigenvalues of S.

The volume of that ellipsoid is equal to |S|^(1/2). Note that |S| = product of all eigenvalues. Note that |S| is the "generalized sample variance", or a measure of the spread of the combination of multiple variables taken as a whole.

If any eigenvalue is zero, the volume "degenerates" to zero.

The term "degenerate" usually implies a trivial special case, or a reduction in dimension to a simpler problem.

Subject: Types of Matrices

Participant's Question:

Is there a document or reference that explains positive definite, positive semi-definite or non-negative definite in more detail?

Instructor's Response:

A "positive definite" symmetric matrix is one with all positive eigenvalues, so x'Ax > 0 for all x other than x = 0.

A "positive semi-definite" symmetric matrix is one for which *at least one* of the eigenvalues is zero, and all the others are positive.

A "non-negative definite" symmetric matrix is one for which all eigenvalues are greater than or equal to zero.

By these definitions, "non-negative definite" includes *both* "positive definite" and "positive semi-definite" matrices. It differs from "positive semi-definite" in requiring at least one eigenvalue be zero. The set of "non-negative definite" symmetric matrices is the union of the sets of "positive definite" and "positive semi-definite" matrices.

If all the eigenvalues greater than zero, the symmetric matrix is both "positive definite" and "non-negative definite". If some of the eigenvalues are zero and the others are greater than zero, the symmetric matrix is both "positive semi-definite" and "non-negative definite". If any of the eigenvalues are negative, it doesn't belong to any of the three classes.

Subject: Use of eigenvalues and eigenvectors

Participant's Question:

What is the use of eigenvalue/eigenvectors in statistical analysis other than the principal compoment analysis?? Are the eigenvalues/vectors also frequently used - used as a opeartor - in say logistic regressions etc?

Instructor's Response:

1. Principal component analysis.
2. Factor analysis.
3. Canonical correlation.

In addition, decomposition of a matrix into the first 1, 2, or 3 largest eigenvalues and associated eigenvectors is a first step in many multivariate methods to reduce dimensionality before proceeding with further analysis or a test. For example, plotting vs. the first two eigenvectors (in an ordering of decreasing eigenvalues) is useful in cluster analysis. If you have 50 variables, principal components are about the only way of visualizing the data. (There are a few nonparametric methods, such as Chernoff faces or star diagrams that try to do something similar, but even then they work better on the principal components.)

Finally, eigenset analysis is used frequently in proving theorems.

Subject: Spectral decomposition

Participant's Question:

Is spectral decomposition the same as diagonalization?

Instructor's Response:

Spectral decomposition is almost the same thing as diagonalization. For symmetric real A, U' A U = D is diagonal (with eigenvalues on the diagonal) if U = (normalized) eigenvector matrix. So A = U D U', as U' is the inverse of U. Thus U "diagonalizes" A. Spectral decomposition says somewhat the same thing, but as a sum formula:

A = d1 e1 x e1' + ... + dn en x en'

where di = i-th eigenvalue and ei = i-th normalized eigenvector.

The two formulas are not immediately and obviously identical, so there is some difference, and a short proof required perhaps.

Where spectral decomposition is most useful as a direct concept is in infinite dimensional linear algebra, where the sum formula becomes an infinite series that represents the action of, say, a differential operator. This is the basis of what is called Stuerm-Liouville theory. It is used to elegantly solve linear differential equations. Fourier series are a subcase, as is the mathematics of quantum mechanics in physics. In numerical analysis, it is called the Galerkin method. Luckily in statistics we only deal with finite dimensions.

Subject: R and S matrices

Participant's Question:

I have a question pertaining to R and S matrices. Could you briefly explain the difference in terms of population model fits. I sometimes get the R or S matrix is singular or R matrix is non-positive semi-definite resulting in covariance step termination. Additionally, what sig can obtaining the eigenvalues and computing a condition number (ratio of largest to smallest eigenvalues) have on identifying ill conditioned models? What information do the eigenvalues provide in relation to the variance-covariance matrix values?

Instructor's Response:

I'm not sure what you mean by R and S matrices. I'm going to assume that you denote the sample covariance matrix by S and the sample correlation matrix by R.

The correlation matrix R is the covariance matrix S with each i,j element divided by the standard deviation of the row i variable and the column j variable. This makes all elements of R between -1 and +1, with 1.0 on the diagonal.
Normally both R and S are positive definite, so long as the variables involved are linearly independent.
If there are one or more constraints among the variables, a corresponding number of eigenvalues will be zero. Normally this is dealt with by reducing the set of variables and trying again. Sometimes missing data can give rise to singularity.
The principal components analysis situation corresponds not to exact dependence, but only approximate dependence. In this case the smallest eigenvalue of R may be 0.0001 or even 1e-8, particularly if there are a lot of variables involved. This is called "approximate collinearity", and happens quite frequently.
The base-10 logarithm of the ratio of the largest eigenvalue to the smallest ("condition number") gives an estimate of the number of decimal places that are lost to linear dependence. If double precision is 15 decimal places, and the condition number is 15 or more, then some numerical algorithms will give nonsense for answers. Note that the log of the ratio of the largest to smallest pivot elements (divisors) used in forming the matrix inverse is an estimate of the condition number.
The eigenvalues of the covariance matrix give the variances of the principal components, or the variances of the uncorrelated (orthogonal) coordinate system represented by the normalized eigenvectors. See the attached PDF for a few more words about this.

Subject: Eigenvalue & Eigenvector Names
Participant's Question:
Here is a "nomenclature" question. Where does the name "eigen" come from? Does it have some meaning or is it associated with ther person's name who identified it? I understand how the value and vector part of the word makes sence, but "eigen" seems sort of strange. Any background?

Instructor's Response:
Eigen" is a German word meaning "same" or "identical", but also "characteristic", "idiosyncratic" or "innate". So you might take it as the German equivalent of "characteristic" used in English. See:
http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors
Words and symbols get attached to things by the first people who proselytize them. Thus "z" for the normal deviate. Sometimes it's very hard to get rid of concepts generated by famous people, even when they are replaced by better concepts later.
A good example of this is hypothesis testing. After it became entrenched, Fisher introduced the concept of confidence (actually, fiducially) interval estimates. This was a much better solution to the problem. Hypothesis tests tell only if an experiment is large enough or well enough designed to detect (aka "significant") an effect. It was very unfortunate that the word "significance" was attached to the process. It would have been much better to have used the term "detectable". This has led to a century of confusion and misunderstanding about what hypothesis testing means, and what a "significant" result means. It would be better to expunge the whole concept from statistics. This is happening now, but glacially slowly.

WEEK 4

Subject: PCA
Participant's Question:
I was curious about the interpretation of the PCA. Can you provide more detail? Let's use the stock price example. Near the end of the explanation, it is written that the three indexes could be "replaced" with a single new index. Can you explain the meaning of the coefficients in more detail, and also the z values? It's said the z value is obtained by "dividing the difference from the mean by the standard deviation of the variable". What "difference"? I'm assuming the standard deviation is the standard deviation of the week end stock index values for each individual index. But the "difference" doesn't make sense to me. What is meant by that? When are we measuring this "difference"? Are we essentially "re-weighting" this index every week, and this "difference" is the residual from the mean over the time period measured and the current end of week value for the index? And then assuming we calculate the z-values and come to a solution, can you interpret that solution?

To be quite frank, I was beginning to wonder what the purpose of all we've been learning was for. And the PCA examples seem to explain some important applications of the concepts we've learned. But I'm a bit unclear on interpreting the results, particularly in this example, since it's something I'm a bit more familiar with given my finance background. So if I can understand it a little more in this context, that would be helpful.
Instructor's Response:
1. The 'z' values are the 'z-scores' learned in elementary statistics. They normalize a variable X by subtracting the mean of X (which moves the mean of z to zero) and dividing by the standard deviation of X (which makes the standard deviation of z equal to 1). z = (X - mean(X))/sd(X). The mean is the mean of the entire dataset for the variable X, and similarly the standard deviation. You could use the median instead of the mean, or the interquartile range instead of the standard deviation, or even constants that you can justify as reasonable values. Sometimes people subtract the lowest value of X (if it is bounded) and divide by the difference between the max and min of X. The object is to remove scale and center from the normalized variables to put all variables on par with each other numerically.

PCA is typically performed on normalized variables, otherwise the large ranging variables will swamp the others just because of units of measure.

2. In the example with DJIA, SP500 and NASDAQ, you might expect the 3 indices measure different things, as the DJIA is based on a small number of very large companies, SP500 is based on a large number (500) of stocks, and NASDAQ is based specifically on technology companies. Instead, PCA indicates they all measure approximately the same thing, indicated by the one large eigenvalue accounting for 83% of the effect. The equation for z_F shows how to compute this new common variable from the normalized values for the 3 indices. The coefficients in this equation are the coefficients of the corresponding eigenvector. The 3 coefficients are approximately equal,

so z_F corresponds roughly to an equally weighted average of the 3 normalized indices. An interesting null hypothesis here would be that the 3 normalized indices are each estimators of the same underlying variable, so the best combined estimate is the arithmetic mean of the 3.

3. To make a new z_F estimate for a new time period, you would have to normalize the new DJIA, SP500 and NASDAQ indices for that time period, then use the equation to calculate z_F. You could plot a stock price vs. z_F, or do a fit vs. z_F, or a variety of other statistically analyses that make sense for the application.

4. If you would like more details on the application, this would be found (for this example) in a course on econometrics or time series analysis. A course in Factor Analysis or Structural Equation Modeling would also explain the underlying models, but are not directly in point to finance.

5. The application examples given in this course are meant to illustrate the type of use you might encounter for the abstract methods discussed. The coverage is necessarily minimal. There is a 'chicken-and-the-egg' problem here: If students take the applied course with first taking matrix algebra, they have trouble understanding what is being done and the notation used. If they take matrix algebra first without the applied course, they learn abstract methods but don't know what they're good for. We've tried to provide a smattering of applied exercises to sustain credibility that there is actually a use for the methodologies learned in this course.

Subject: Spectral decomposition
Participant's Question:
I am not quite getting the concept of spectral decomposition. Can you please elaborate it?
Instructor's Response:
The 'spectral decomposition' of a symmetric matrix is the fact that it diagonalizable into the form
(1) **A** = **E D E**'
where **A** is the symmetric matrix, **D** is a diagonal matrix with **A**'s eigenvalues on the diagonal, and **E** is the normalized eigenvector matrix (each column is an eigenvector).
Eq.(1) is frequently given as the 'spectral decomposition' of **A**.
Eq.(1) follows from the vectorized eigenvector equation
(2) **A E** = **E D** (generalization of **A e** = lambda **e**)
Now, if you take into account **E**'s column orthogonality, if you work out eq.(1), you get
(3) **A** = sum(1:n) { lambda(i) **e**(i) **e**(i)' }
where each **e**(i) **e**(i)' is a matrix.
The term 'spectral' relates to the eigenvalues, which are called the 'spectrum' of **A**.
'Decomposition' relates to breaking down **A** into n independent components that sum to it.
The most obvious direct application to statistics is in Principal Components Analysis.
Let me know if this satisfies you, or if you would like some type of numerical example.

Subject: Homogeneous system of equations

Participant's Question:
Where exactly would I encounter homogeneous system of equations in Statistics?
Instructor's Response:
Homogeneous equations of the form A x = 0 come up from time to time.
An example from our course is (A - lambda I) x = 0, the eigenvalue equation.
Another example is with 'contrasts' in ANOVA. There typically the sum of levels of a factor are constrained to sum to zero, so linear combination equations involved them are homogeneous and rank less than n.
Also, we cover homogeneous equations in this course because the general solution of an inhomogeneous system of linear equations is a particular inhomogeneous solution plus the general solution of the homogeneous equations. (The same principle is used to solve inhomogeneous linear differential equations with constant coefficients.)
Homogeneous equations arise naturally in systems with rank less than n.
As you appear to be very interested in linear algebra, you should note that the theory extends to continuous as well as discrete values of the subscripts. Vectors then become functions and matrices become integral operators. (This is why differential equations are solved the same way as matrix equations.) Certain ordinary differential equations are called 'Sturm-Liouville' problems, and have eigenvalues and associated eigenfunctions. These functions include Fourier series, orthogonal polynomials (Legendre, Chebyshev, Laguerre, Hermite, etc.) and many others. This is a general way of solving linear operator equations in applied mathematics. A rich field. It is also the basis of 'quantum mechanics' in chemical-physics. (You might say that matrix algebra was the basis of 19th and 20th century physics.)

Subject: PCA on covariance and correlation
Participant's Question: Principal Component Analysis on **Covariance** and Principal Component Analysis on **Correlation**,what is the difference? (The answers/eigensets are different.) And when do you use each method?

Instructor's Response: PCA can be done on any symmetric, positive-definite matrix, which includes both the covariance and correlation matrices.

Which matrix is of interest to a statistician depends upon the context of the problem and subject-matter expertise. The question is 'Does the scale of variables matter?'

Normally the unit of measure of a variable is arbitrary. The standard deviation is in this unit as well, so the variance scales with the square of the unit. This is disinformative and an annoyance. For example, suppose a variable is 'Weight'. The experimenter could equally choose lb, kg, g, mg, etc. If you change the unit of measure, all numbers are changed by the corresponding conversion factor. E.g., if the original recording were in kg, then changing to g would multiply every number by 1000. If you did PCA on the covariance matrix, then the 'Weight' variable would be very important (large variance) if measured in g, but perhaps not if measured in kg. For most problems you want the inference to be scale and location invariant, so you scale and re-zero the variables (e.g., subtract mean and divided by standard deviation) before doing multivariate analysis

(including model fits where you want to interpret coefficients). Thus the common choice is to do PCA on the correlation matrix and not the covariance matrix for this reason.

On the other hand, sometimes the data are 'pre-scaled' a priori. E.g., you are a psychologist investigating different IQ tests. You give a battery of 8 different tests to the same subjects. Each IQ test has been pre-scaled by design to have a supposed mean of 100 and a standard deviation of 15 (based on its original validation experiment). Now the units of measurement of the 8 tests are actually directly comparable to each other. (I.e., a 120 IQ on one test compares directly to an IQ of 118 from another test, if done on the same subject). Therefore you don't need to do a posteriori scaling. PCA would be done on the covariance matrix. Similarly, if you are fitting a model for prediction purposes or if the coefficients are most interpretable in the original units, you would fit a linear model in the original units. Another example would by survey question answers, all on the same scale 1 to 10.

INDEX