A Introduction to Matrix Algebra and Principal Components Analysis

PSYC 943 (930): Fundamentals
of Multivariate Modeling

Lecture 10: September 28, 2012

Today's Class

- An introduction to matrix algebra
  - Scalars, vectors, and matrices
  - Basic matrix operations
  - Advanced matrix operations

- An introduction to principal components analysis

- An introduction to SAS PROC IML
  - Interactive matrix language

Why Learning a Little Matrix Algebra is Important

- Matrix algebra is the alphabet of the language of statistics
  - ➢ You will most likely encounter formulae with matrices very quickly

- For example, imagine you were interested in analyzing some repeated measures data...but things don't go as planned
  - ➢ From the SAS User's Guide (PROC MIXED):

Formulation of the Mixed Model

The previous general linear model is certainly a useful one (Sea
although you still assume normality.

The mixed model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where everything is the same as in the general linear model exc
Henderson (1990) and Searle, Casella, and McCulloch (1992) fc

A key assumption in the foregoing analysis is that $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are

$$E\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

$$Var\begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

The variance of $\mathbf{y}$ is, therefore, $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. You can model

Estimating Covariance Parameters in the Mixed Model

Estimation is more difficult in the mixed model than in the general linear model. Not only do y

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

However, it requires knowledge of $\mathbf{V}$ and, therefore, knowledge of $\mathbf{G}$ and $\mathbf{R}$. Lacking such info

In many situations, the best approach is to use *likelihood-based* methods, exploiting the ass
(REML). A favorable theoretical property of ML and REML is that they accommodate data tha

PROC MIXED constructs an objective function associated with ML or REML and maximizes i

$$ML: \quad l(\mathbf{G},\mathbf{R}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{n}{2}\log(2\pi)$$

$$REML: \quad l_R(\mathbf{G},\mathbf{R}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{n-p}{2}\log(2\pi)\}$$

where $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and p is the rank of $\mathbf{X}$. PROC MIXED actually minimiz
analytical details for implementing a QR-decomposition approach to the problem. Wolfinger, T

Introduction and Motivation

- Nearly all multivariate statistical techniques are described with matrix algebra

- When new methods are developed, the first published work typically involves matrices
  - It makes technical writing more concise – formulae are smaller

- Have you seen:
  - $(X^TX)^{-1}X^Ty$
  - $\mathbf{\Lambda\Phi\Lambda}^T + \mathbf{\Psi}$

- **Useful tip:** matrix algebra is a great way to get out of conversations and other awkward moments

Definitions

- We begin this class with some general definitions (from dictionary.com):

    - **Matrix:**
        1. A rectangular array of numeric or algebraic quantities subject to mathematical operations
        2. The substrate on or within which a fungus grows

    - **Algebra:**
        1. A branch of mathematics in which symbols, usually **letters** of the alphabet, represent numbers or members of a specified set and are used to represent quantities and to express general relationships that hold for all members of the set

        2. A set together with a pair of **binary operations** defined on the set. Usually, the set and the operations include an **identity element**, and the operations are **commutative** or **associative**

Why Learn Matrix Algebra

- Matrix algebra can seem very abstract from the purposes of this class (and statistics in general)

- Learning matrix algebra is important for:
  - Understanding how statistical methods work
    - And when to use them (or not use them)
  - Understanding what statistical methods mean
  - Reading and writing results from new statistical methods

- Today's class is the first lecture of learning the language of multivariate statistics

DATA EXAMPLE AND SAS

A Guiding Example

- To demonstrate matrix algebra, we will make use of data

- Imagine that somehow I collected data SAT test scores for both the Math (SATM) and Verbal (SATV) sections of 1,000 students

- The descriptive statistics of this data set are given below:

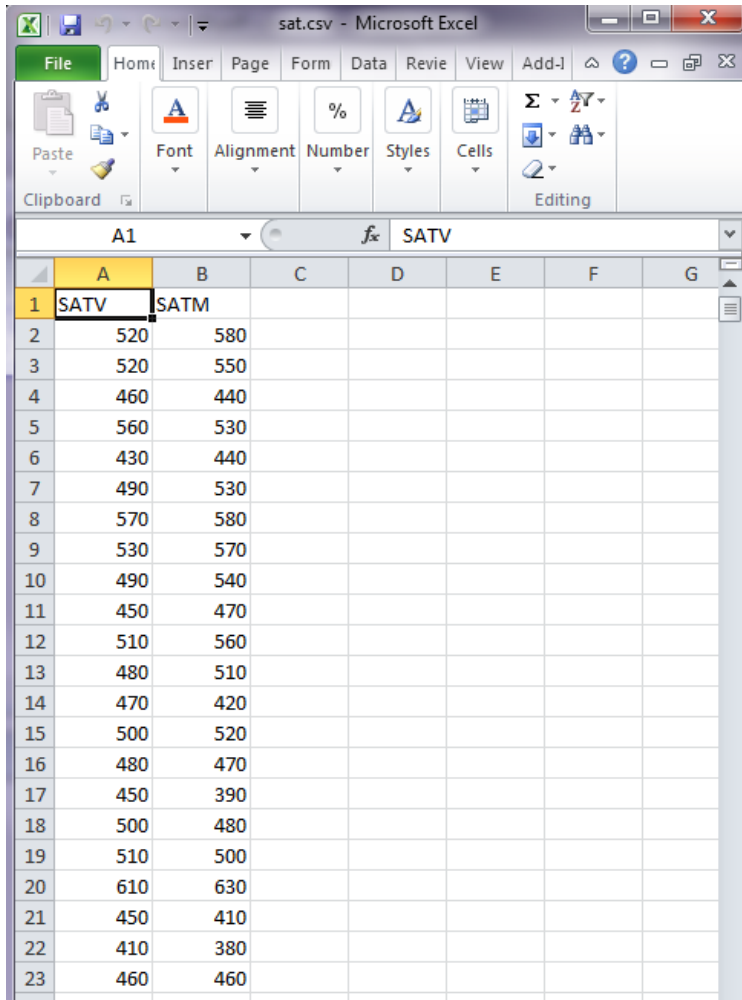| Statistic	SATV	SATM
Mean	499.3	498.3
SD	49.8	81.2

Correlation		
SATV	1.00	0.78
SATM	0.78	1.00

The Data…

In Excel:

In SAS:

Matrix Computing: PROC IML

- To help demonstrate the topics we will discuss today, I will be showing examples in SAS PROC IML

- The Interactive Matrix Language (IML) is a scientific computing package in SAS that typically used for statistical routines that aren't programed elsewhere in SAS

- Useful documentation for IML:

http://support.sas.com/documentation/cdl/en/imlug/64248/HTML/default/viewer.htm#langref_toc.htm

- A great web reference for IML:

http://www.psych.yorku.ca/lab/sas/iml.htm

PROC IML Basics

- Proc IML is a proc step in SAS that runs without needing to use a preliminary data step

- To use IML the following lines of syntax are placed in a SAS file:

```
PROC IML;
  RESET PRINT;

  *IML SYNTAX GOES IN HERE;

QUIT;
```

- The "reset print;" line makes every result get printed in the output window

- The IML syntax will go between the "reset print;" and the "quit;"
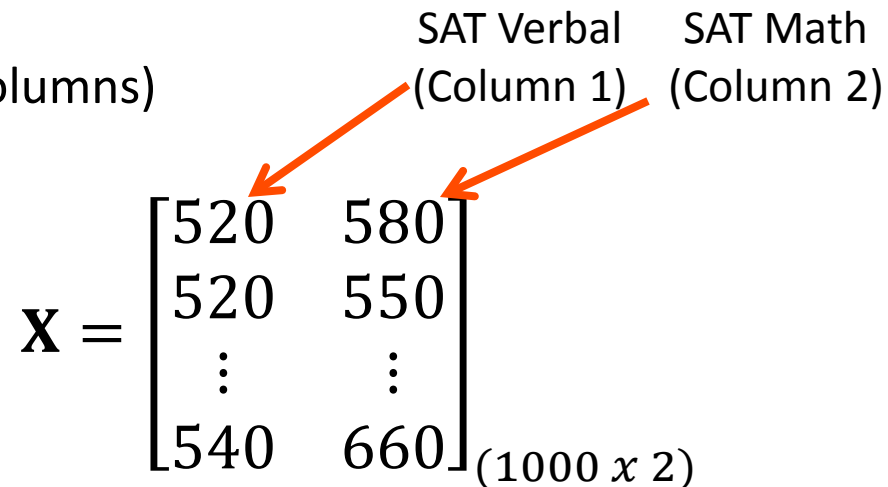
DEFINITIONS OF MATRICES, VECTORS, AND SCALARS

Matrices

- A matrix is a rectangular array of data
  - Used for storing numbers

- Matrices can have unlimited dimensions
  - For our purposes all matrices will have two dimensions:
    - Row
    - Columns

- Matrices are symbolized by **boldface** font in text, typically with capital letters
  - Size (r rows x c columns)

SAT Verbal   SAT Math
(Column 1)   (Column 2)

$$\mathbf{X} = \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix}_{(1000 \, x \, 2)}$$

Vectors

- A vector is a matrix where one dimension is equal to size 1
  - Column vector: a matrix of size $r \times 1$

$$\boldsymbol{x}_{\cdot 1} = \begin{bmatrix} 520 \\ 520 \\ \vdots \\ 540 \end{bmatrix}_{1000 \times 1}$$

  - Row vector: a matrix of size $1 \times c$

$$\boldsymbol{x}_{1\cdot} = \begin{bmatrix} 520 & 580 \end{bmatrix}_{1 \times 2}$$

- Vectors are typically written in **boldface** font text, usually with lowercase letters

- The dots in the subscripts $\boldsymbol{x}_{\cdot 1}$ and $\boldsymbol{x}_{1\cdot}$ represent the dimension aggregated across in the vector
  - $\boldsymbol{x}_{1\cdot}$ is the first row and **all** columns of $\mathbf{X}$
  - $\boldsymbol{x}_{\cdot 1}$ is the first column and **all** rows of $\mathbf{X}$
  - Sometimes the rows and columns are separated by a comma (making it possible to read double-digits in either dimension)

Matrix Elements

- A matrix (or vector) is composed of a set of elements
  - Each element is denoted by its position in the matrix (row and column)

- For our matrix of data **X** (size 1000 rows and 2 columns), each element is denoted by:

$$x_{ij}$$

  - The first subscript is the index for the rows: i = 1,…,r (= 1000)
  - The second subscript is the index for the columns: j = 1,…,c (= 2)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{1000,1} & x_{1000,2} \end{bmatrix}_{(1000\ x\ 2)}$$

Scalars

- A scalar is just a single number

- The name scalar is important: the number "scales" a vector – it can make a vector "longer" or "shorter"

- Scalars are typically written without boldface:
$$x_{11} = 520$$

- Each element of a matrix is a scalar

Matrix Transpose

- The transpose of a matrix is a reorganization of the matrix by switching the indices for the rows and columns

$$\mathbf{X} = \begin{bmatrix} 520 & 580 \\ 520 & 550 \\ \vdots & \vdots \\ 540 & 660 \end{bmatrix}_{(1000\ x\ 2)}$$

$$\mathbf{X}^T = \begin{bmatrix} 520 & 520 & \cdots & 540 \\ 580 & 550 & \cdots & 660 \end{bmatrix}_{(2\ x\ 1000)}$$

- An element x_{ij} in the original matrix $\mathbf{X}$ is now x_{ji} in the transposed matrix $\mathbf{X}^T$

- **Transposes are used to align matrices for operations where the sizes of matrices matter (such as matrix multiplication)**

Types of Matrices

- **<u>Square Matrix:</u>** A square matrix has the same number of rows and columns

  ➢ Correlation/covariance matrices are square matrices

- **<u>Diagonal Matrix:</u>** A diagonal matrix is a square matrix with non-zero diagonal elements ($x_{ij} \neq 0$ for $i = j$) and zeros on the off-diagonal elements ($x_{ij} = 0$ for $i \neq j$):

$$\mathbf{A} = \begin{bmatrix} 2.759 & 0 & 0 \\ 0 & 1.643 & 0 \\ 0 & 0 & 0.879 \end{bmatrix}$$

  ➢ We will use diagonal matrices to form correlation matrices

- **<u>Symmetric Matrix:</u>** A symmetric matrix is a square matrix where all elements are reflected across the diagonal ($a_{ij} = a_{ji}$)

  ➢ Correlation and covariance matrices are symmetric matrices

VECTORS

Vectors in Space…

- Vectors (row or column) can be represented as lines on a Cartesian coordinate system (a graph)

- Consider the vectors: $\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

- A graph of these vectors would be:
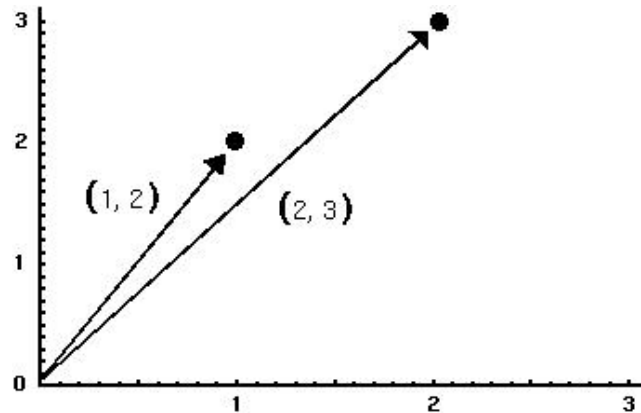
- **Question:** how would a column vector for each of our example variables (SATM and SATV) be plotted?

Vector Length

- The length of a vector emanating from the origin is given by the Pythagorean formula
  - This is also called the Euclidean distance between the endpoint of the vector and the origin

$$L_{\mathbf{x}} = \sqrt{x_{11}^2 + x_{21}^2 + \cdots + x_{r1}^2} = \|\mathbf{x}\|$$

- From the last slide: $\|\mathbf{a}\| = \sqrt{5} = 2.24$; $\|\mathbf{b}\| = \sqrt{13} = 3.61$

- From our data:
$$\|\mathbf{SATV}\| = 15{,}868.138; \|\mathbf{SATM}\| = 15{,}964.42$$

- **In data:** length is an analog to the standard deviation
  - In mean-centered variables, the length is the square root of the sum of mean deviations (not quite the SD, but close)

Vector Addition

- Vectors can be added together so that a new vector is formed

- Vector addition is done element-wise, by adding each of the respective elements together:
  - The new vector has the same number of rows and columns

$$\mathbf{c} = \mathbf{a} + \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

  - Geometrically, this creates a new vector along either of the previous two
    - Starting at the origin and ending at a new point in space

- **In Data:** a new variable (say, SAT total) is the result of vector addition

$$SAT_{TOTAL} = x_{.1} + x_{.2}$$

Vector Addition: Geometrically
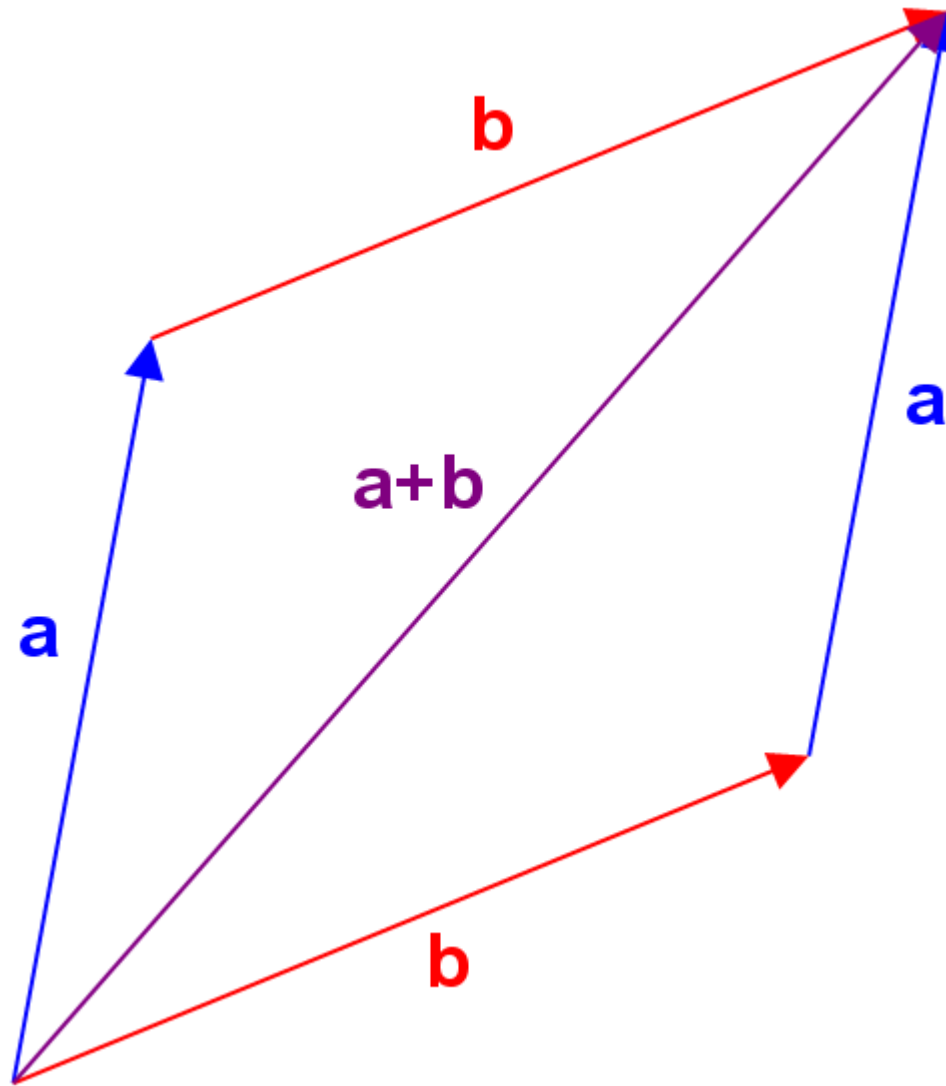
Vector Multiplication by Scalar

- Vectors can be multiplied by scalars
  - All elements are multiplied by the scalar

$$\mathbf{d} = 2\mathbf{a} = 2\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

- Scalar multiplication changes the length of the vector:

$$\|\mathbf{d}\| = \sqrt{2^2 + 4^2} = \sqrt{20} = 4.47$$

- This is where the term scalar comes from: a scalar ends up "rescaling" (resizing) a vector

- **In Data:** the GLM (where **X** is a matrix of data) the fixed effects (slopes) are scalars multiplying the data
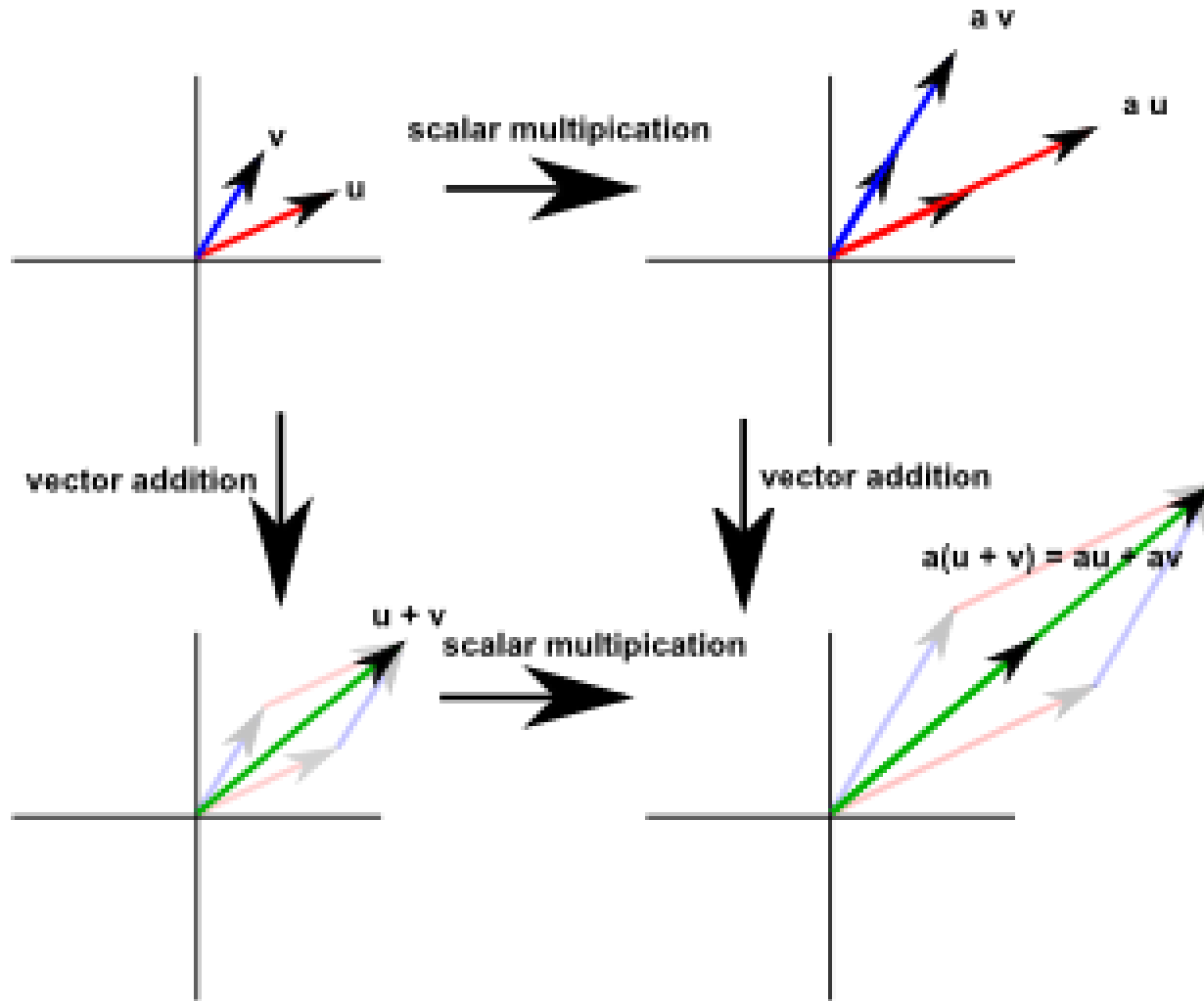
Linear Combinations

- Addition of a set of vectors (all multiplied by scalars) is called a linear combination:

$$\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$$

- Here, $\mathbf{y}$ is the linear combination
  - For all k vectors, the set of all possible linear combinations is called their span
  - Typically not thought of in most analyses – but when working with things that don't exist (latent variables) becomes somewhat important

- **In Data**: linear combinations happen frequently:
  - Linear models (i.e., Regression and ANOVA)
  - Principal components analysis (later today)

Linear Dependencies

- A set of vectors are said to be linearly dependent if

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k = 0$$

-and-

$$a_1, a_2, \ldots, a_k \text{ are all } \mathbf{not} \text{ zero}$$

- Example: let's make a new variable – SAT Total:

$$\mathbf{SAT_{total}} = 1 * \mathbf{SATV} + 1 * \mathbf{SATM}$$

- The new variable is linearly dependent with the others:

$$1 * \mathbf{SATV} + 1 * \mathbf{SATM} - 1 * \mathbf{SAT_{total}} = \mathbf{0}$$

- **In Data**: (multi) collinearity is a linear dependency. Linear dependencies are bad for statistical analyses that use matrix inverses (discussed soon).

Normalizing Vectors

- Normalizing vectors is the process of rescaling a vector (multiplication by scalar) so that the length of the vector is equal to one:

$$\mathbf{x}_{norm} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$$

- From our example vector:

$$\mathbf{a}_{norm} = \frac{\mathbf{a}}{\|\mathbf{a}\|} = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sqrt{5}} \\ \dfrac{2}{\sqrt{5}} \end{bmatrix}$$

- As such, the length of the vector is:

$$\|\mathbf{a}\| = \sqrt{\left(\frac{1}{\sqrt{5}}\right)^2 + \left(\frac{2}{\sqrt{5}}\right)^2} = 1$$

Normalizing in Statistics

- We will encounter normalized vectors later today – in principal components analysis

- Generally speaking, entities that have no real space (such as principal components and "latent" factors) have corresponding vectors that are normalized
  - ➢ Normalization allows for a solution to be found
  - ➢ It's an arbitrary standard (but one that is well agreed upon)

Inner (Dot) Product of Vectors

- An important concept in vector geometry is that of the inner product of two vectors
  - The inner product is also called the dot product

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = a_{11}b_{11} + a_{21}b_{21} + \cdots + a_{N1}b_{N1} = \sum_{i=1}^{N} a_{i1}b_{i1}$$

- The dot or inner product is related to the angle between vectors and to the projection of one vector onto another

- From our example: $\mathbf{a} \cdot \mathbf{b} = 1 * 2 + 2 * 3 = 8$
- From our data: $\boldsymbol{x}_{.1} \cdot \boldsymbol{x}_{.2} = 251{,}928{,}400$

- **In data:** the angle between vectors is related to the correlation between variables and the projection is related to regression/ANOVA/linear models

Angle Between Vectors

- As vectors are conceptualized geometrically, the angle between two vectors can be calculated

$$\theta_{ab} = \cos^{-1}\left(\frac{\mathbf{a} \cdot \boldsymbol{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}\right)$$

- From the example:

$$\theta_{ab} = \cos^{-1}\left(\frac{8}{\sqrt{5}\sqrt{13}}\right) = 0.12$$

- From our data:

$$\theta_{SATV,SATM} = \cos^{-1}\left(\frac{251,928,400}{\sqrt{15,868.138}\sqrt{15,946.42}}\right) = 0.105$$

In Data: Cosine Angle = Correlation

- If you have data that are:
  - Placed into vectors
  - Centered by the mean (subtract the mean from each observation)

- ...then the cosine of the angle between those vectors is the correlation between the variables:

$$r_{ab} = cos(\theta_{ab}) = \frac{\mathbf{a} \cdot \boldsymbol{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{N}(a_{i1} - \bar{a})(b_{i1} - \bar{b})}{\sqrt{\sum_{i=1}^{N}(a_{i1} - \bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_{i1} - \bar{b})^2}}$$

For the SAT example data (using mean centered variables):

$$r_{SATV,SATM} = \cos(\theta_{SATVc,SATMc}) = \cos\left(\frac{3,132,223.6}{1,573.956 * 2,567.0425}\right)$$
$$= .775$$

Vector Projections

- A final vector property that shows up in statistical terms frequently is that of a projection

- The **projection** of a vector **a** onto **b** is the orthogonal projection of **a** onto the straight line defined by **b**
  - ➢ The projection is the "shadow" of one vector onto the other:
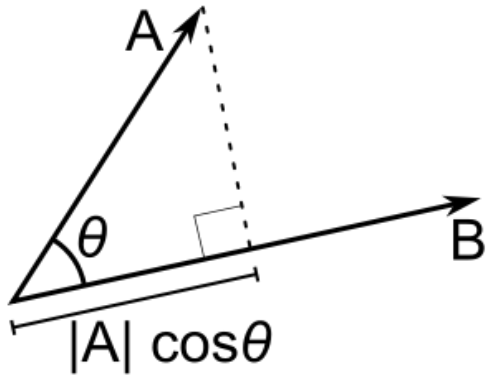
$$\mathbf{a}_{\text{proj } \mathbf{b}} = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{b}\|^2} \mathbf{b}$$
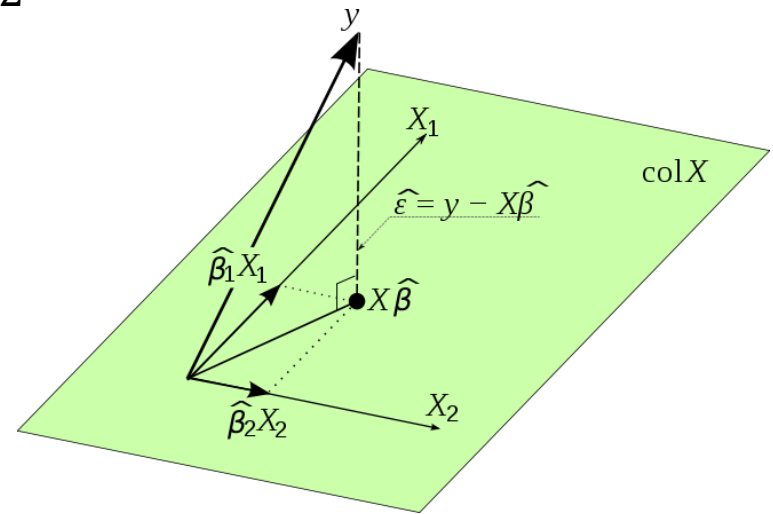
- **In data:** linear models can be thought of as projections

- To provide a bit more context for vector projections, let's consider the projection of SATV onto SATM:

$$SATV_{\text{proj } SATM} = \frac{SATV \cdot SATM}{\|SATM\|^2} SATM$$

- The first portion turns out to be:

$$\frac{SATV \cdot SATM}{\|SATM\|^2} = \frac{251{,}928{,}400}{\|15{,}964.42\|^2} = .475$$

- This is also the regression slope β_1:

$$SATV_p = \beta_0 + \beta_1 SATM_p + e_p$$

```
The GLM Procedure

Dependent Variable: satv


Parameter          Estimate

Intercept     262.4819995
satm            0.4753206
```
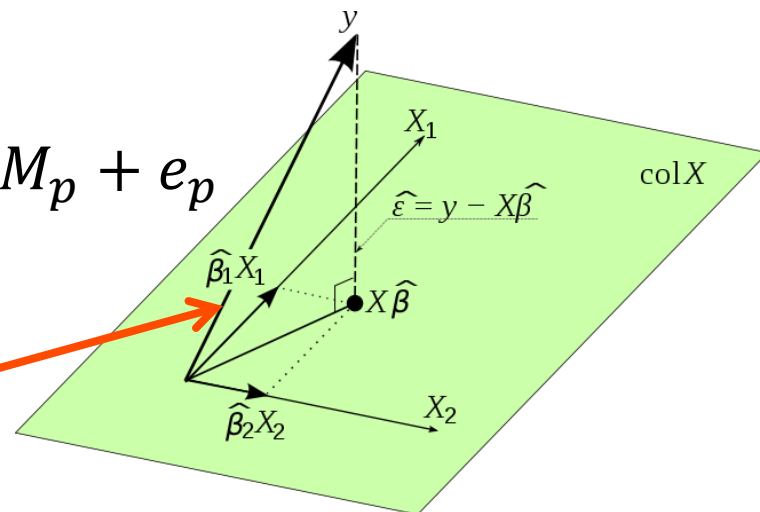
MATRIX ALGEBRA

Moving from Vectors to Matrices

- A matrix can be thought of as a collection of vectors
  - Matrix operations are vector operations on steroids

- Matrix algebra defines a set of operations and entities on matrices
  - I will present a version meant to mirror your previous algebra experiences

- Definitions:
  - Identity matrix
  - Zero vector
  - Ones vector

- Basic Operations:
  - Addition
  - Subtraction
  - Multiplication
  - "Division"

Matrix Addition and Subtraction

- Matrix addition and subtraction are much like vector addition/subtraction

- Rules:
  - Matrices must be the same size (rows and columns)

- Method:
  - The new matrix is constructed of element-by-element addition/subtraction of the previous matrices

- Order:
  - The order of the matrices (pre- and post-) does not matter

Matrix Addition/Subtraction

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \\ a_{31} + b_{31} & a_{32} + b_{32} \\ a_{41} + b_{41} & a_{42} + b_{42} \end{bmatrix} \qquad \mathbf{A} - \mathbf{B} = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} \\ a_{21} - b_{21} & a_{22} - b_{22} \\ a_{31} - b_{31} & a_{32} - b_{32} \\ a_{41} - b_{41} & a_{42} - b_{42} \end{bmatrix}$$

Matrix Multiplication

- Matrix multiplication is a bit more complicated
  - ➢ The new matrix may be a different size from either of the two multiplying matrices

$$\mathbf{A}_{(r \; x \; c)} \mathbf{B}_{(c \; x \; k)} = \mathbf{C}_{(r \; x \; k)}$$

- Rules:
  - ➢ Pre-multiplying matrix must have number of columns equal to the number of rows of the post-multiplying matrix

- Method:
  - ➢ The elements of the new matrix consist of the inner (dot) product of the row vectors of the pre-multiplying matrix and the column vectors of the post-multiplying matrix

- Order:
  - ➢ The order of the matrices (pre- and post-) matters

Matrix Multiplication

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \\ a_{31}b_{11} + a_{32}b_{21} & a_{31}b_{12} + a_{32}b_{22} & a_{31}b_{13} + a_{32}b_{23} \\ a_{41}b_{11} + a_{42}b_{21} & a_{41}b_{12} + a_{42}b_{22} & a_{41}b_{13} + a_{42}b_{23} \end{bmatrix}$$

Multiplication in Statistics

- Many statistical formulae with summation can be re-expressed with matrices

- A common matrix multiplication form is: $\mathbf{X}^T\mathbf{X}$
  - Diagonal elements: $\sum_{i=1}^{N} X_i^2$
  - Off-diagonal elements: $\sum_{i=1}^{N} X_{ia}X_{ib}$

- For our SAT example:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{N} SATV_i^2 & \sum_{i=1}^{N} SATV_i SATM_i \\ \sum_{i=1}^{N} SATV_i SATM_i & \sum_{i=1}^{N} SATM_i^2 \end{bmatrix}$$

$$= \begin{bmatrix} 251{,}797{,}800 & 251{,}928{,}400 \\ 251{,}928{,}400 & 254{,}862{,}700 \end{bmatrix}$$

Identity Matrix

- The identity matrix is a matrix that, when pre- or post- multiplied by another matrix results in the original matrix:

$$\mathbf{AI} = \mathbf{A}$$
$$\mathbf{IA} = \mathbf{A}$$

- The identity matrix is a square matrix that has:
  - Diagonal elements = 1
  - Off-diagonal elements = 0

$$I_{(3\ x\ 3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Zero Vector

- The zero vector is a column vector of zeros

$$\mathbf{0}_{(3 \; x \; 1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

- When pre- or post- multiplied the result is the zero vector:

$$\mathbf{A0} = \mathbf{0}$$
$$\mathbf{0A} = \mathbf{0}$$

Ones Vector

- A ones vector is a column vector of 1s:

$$\mathbf{1}_{(3 \times 1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

- The ones vector is useful for calculating statistical terms, such as the mean vector and the covariance matrix

Matrix "Division": The Inverse Matrix

- Division from algebra:
  - First: $\dfrac{a}{b} = \dfrac{1}{b} a = b^{-1} a$
  - Second: $\dfrac{a}{a} = 1$

- "Division" in matrices serves a similar role
  - For square and symmetric matrices, an inverse matrix is a matrix that when pre- or post- multiplied with another matrix produces the identity matrix:
  $$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$
  $$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

- Calculation of the matrix inverse is complicated
  - Even computers have a tough time

- Not all matrices can be inverted
  - Non-invertible matrices are called singular matrices
    - In statistics, singular matrices are commonly caused by linear dependencies

The Inverse

- **<u>In data:</u>** the inverse shows up constantly in statistics
  - Models which assume some type of (multivariate) normality need an inverse covariance matrix

- Using our SAT example
  - Our data matrix was size (1000 x 2), which is not invertible
  - However $\mathbf{X}^T\mathbf{X}$ was size (2 x 2) – square, and symmetric

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 251,797,800 & 251,928,400 \\ 251,928,400 & 254,862,700 \end{bmatrix}$$

  - The inverse is:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 3.61E-7 & -3.57E-7 \\ -3.57E-7 & 3.56E-7 \end{bmatrix}$$

- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} =$
  $$\mathbf{A} + (\mathbf{B} + \mathbf{C})$$

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

- $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$

- $(c + d)\mathbf{A} = c\mathbf{A} + d\mathbf{A}$

- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$

- $(cd)\mathbf{A} = c(d\mathbf{A})$

- $(c\mathbf{A})^T = c\mathbf{A}^T$

- $c(\mathbf{AB}) = (c\mathbf{A})\mathbf{B}$

- $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$

- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$

- $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

- For x_j such that Ax_j exists:

$$\sum_{j=1}^{N} \mathbf{A}\mathbf{x}_j = \mathbf{A}\sum_{j=1}^{N} \mathbf{x}_j$$

$$\sum_{j=1}^{N} (\mathbf{A}\mathbf{x}_j)(\mathbf{A}\mathbf{x}_j)^T =$$

$$\mathbf{A}\left(\sum_{j=1}^{N} \mathbf{x}_j \mathbf{x}_j^T\right)\mathbf{A}^T$$

ADVANCED MATRIX OPERATIONS

Advanced Matrix Functions/Operations

- We end our matrix discussion with some advanced topics
  - All related to multivariate statistical analysis

- To help us throughout, let's consider the correlation matrix of our SAT data:

$$\mathbf{R} = \begin{bmatrix} 1.00 & 0.78 \\ 0.78 & 1.00 \end{bmatrix}$$

Matrix Trace

- For a square matrix $\mathbf{A}$ with p rows/columns, the trace is the sum of the diagonal elements:

$$tr\mathbf{A} = \sum_{i=1}^{p} a_{ii}$$

- For our data, the trace of the correlation matrix is 2
  - ➢ For all correlation matrices, the trace is equal to the number of variables because all diagonal elements are 1

- The trace is considered the total variance in multivariate statistics
  - ➢ Used as a target to recover when applying statistical models

Matrix Determinants

- A square matrix can be characterized by a scalar value called a determinant:

$$\det \mathbf{A} = |\mathbf{A}|$$

- Calculation of the determinant is tedious
  - Our determinant was 0.3916

- The determinant is useful in statistics:
  - Shows up in multivariate statistical distributions
  - Is a measure of "generalized" variance of multiple variables

- If the determinant is positive, the matrix is called **positive definite**
  - Is invertible

- If the determinant is not positive, the matrix is called **non-positive definite**
  - Not invertible

Matrix Orthogonality

- A square matrix $\mathbf{A}$ is said to be orthogonal if:
$$\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$$

- Orthogonal matrices are characterized by two properties:
    1. The dot product of all row vector multiples is the zero vector
        - Meaning vectors are orthogonal (or uncorrelated)
    2. For each row vector, the sum of all elements is one
        - Meaning vectors are "normalized"

- The matrix above is also called **orthonormal**

- Orthonormal matrices are used in principal components

Eigenvalues and Eigenvectors

- A square matrix $\mathbf{A}$ can be decomposed into a set of eigenvalues $\boldsymbol{\lambda}$ and a set of eigenvectors $\mathbf{e}$

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$$

- Each eigenvalue has a corresponding eigenvector
  - The number equal to the number of rows/columns of $\mathbf{A}$
  - The eigenvectors are all orthogonal

- In many classical multivariate statistics, eigenvalues and eigenvectors are used very frequently
  - We will see their use in principal components shortly

Eigenvalues and Eigenvectors Example

- In our SAT example, the two eigenvalues obtained were:

$$\lambda_1 = 1.78$$
$$\lambda_2 = 0.22$$

- The two eigenvectors obtained were:

$$\mathbf{e}_1 = \begin{bmatrix} 0.71 \\ 0.71 \end{bmatrix}; \mathbf{e}_2 = \begin{bmatrix} 0.71 \\ -0.71 \end{bmatrix}$$

- These terms will have much greater meaning in one moment (principal components analysis)

Spectral Decomposition

- Using the eigenvalues and eigenvectors, we can reconstruct the original matrix using a spectral decomposition:

$$\mathbf{A} = \sum_{v=1}^{V} \lambda_v \mathbf{e}_v \mathbf{e}_v^T$$

  ➢ Here v is the index for variable (with a total of V variables)

- For our example, we can get back to our original matrix:

$$\mathbf{R}_1 = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T = 1.78 \begin{bmatrix} .71 \\ .71 \end{bmatrix} [.71 \quad .71] = \begin{bmatrix} .89 & .89 \\ .89 & .89 \end{bmatrix}$$

$$\mathbf{R}_2 = \mathbf{R}_1 + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T = \begin{bmatrix} .89 & .89 \\ .89 & .89 \end{bmatrix} + 0.22 \begin{bmatrix} .71 \\ -.71 \end{bmatrix} [.71 \quad -.71]$$

$$= \begin{bmatrix} 1.00 & 0.78 \\ 0.78 & 1.00 \end{bmatrix}$$

Additional Eigenvalue Properties

- The matrix trace can be found by the sum of the eigenvalues:

$$tr\mathbf{A} = \sum_{v=1}^{V} \lambda_v$$

  ➢ In our example, the $tr\mathbf{R} = 1.78 + .22 = 2$

- The matrix determinant can be found by the product of the eigenvalues

$$|\mathbf{A}| = \prod_{v=1}^{V} \lambda_v$$

  ➢ In our example $|\mathbf{R}| = 1.78 * .22 = .3916$

AN INTRODUCTION TO PRINCIPAL COMPONENTS ANALYSIS

PCA Overview

- Principal Components Analysis (PCA) is a method for re-expressing the covariance (correlation) between a set of variables

  - The re-expression comes from creating a set of new variables (linear combinations) of the original variables

- PCA has two objectives:

  1. Data reduction

     - Moving from many original variables down to a few "components"

  2. Interpretation

     - Determining which original variables contribute most to the new "components"

- My objective: use PCA to demonstrate eigenvalues and eigenvectors

Goals of PCA

- The goal of PCA is to find a set of k principal components (composite variables) that:

  ➢ Is much smaller in number than the original set of V variables

  ➢ Accounts for nearly all of the total variance

    ◆ Total variance = trace of covariance/correlation matrix

- If these two goals can be accomplished, then the set of k principal components contains almost as much information as the original V variables

  ➢ Meaning – the components can now replace the original variables in any subsequent analyses

PCA Features

- PCA often reveals relationships between variables that were not previously suspected
  - New interpretations of data and variables often stem from PCA

- PCA usually serves as more of a means to an end rather than an end it itself
  - Components (the new variables) are often used in other statistical techniques
    - Multiple regression
    - Cluster analysis

- PCA is often intermixed with Exploratory Factor Analysis
  - Don't. Please don't. Please help make it stop.

PCA Details

- Notation: let Y denote our new principal components and let $\mathbf{X}$ be our original data matrix (with N people and V variables)
  - ➢ We will let p be our index for a person

- The new components are linear combinations:

$$Y_{p1} = \mathbf{e}_1^T \mathbf{X} = e_{11} X_{p1} + e_{21} X_{p2} + \cdots + e_{V1} X_{pV}$$
$$Y_{p2} = \mathbf{e}_2^T \mathbf{X} = e_{12} X_{p1} + e_{22} X_{p2} + \cdots + e_{V2} X_{pV}$$
$$\vdots$$
$$Y_{pV} = \mathbf{e}_V^T \mathbf{X} = e_{1V} X_{p1} + e_{2V} X_{p2} + \cdots + e_{VV} X_{pV}$$

- The weights of the components (e_{vk}) come from the eigenvectors of the covariance or correlation matrix
 - ➢ Each variable v has a weight for each component k

Details About the Components

- The components (Y) are formed by the weights of the eigenvectors of the covariance or correlation matrix of the original data
  - The variance of a component is given by the eigenvalue associated with the eigenvector for the component

- Using the eigenvalue and eigenvectors means:
  - Each successive component has lower variance
    - Var(Y$_1$) > Var(Y$_2$) > … > Var(Y$_p$)
  - All components are uncorrelated
  - The sum of the variances of the principal components is equal to the total variance:

$$\sum_{v=1}^{V} Var(Y_v) = tr\mathbf{R} = \sum_{v=1}^{V} \lambda_v$$

PCA on our Example

- We will now conduct a PCA on the correlation matrix of our sample data
  - ➢ This example is given for demonstration purposes – typically we will not do PCA on small numbers of variables
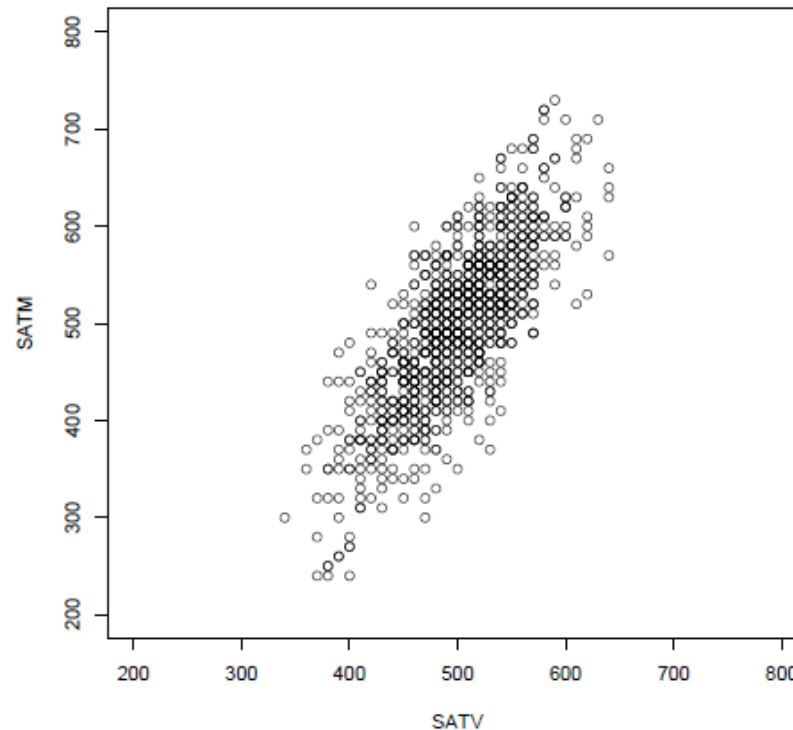
PCA in SAS

- The SAS procedure that does principal components is called proc princomp:

```
proc princomp data=sat out=pcsat_cor;
var satv satm;
run;
```

- The results (look familiar?):

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.77522381	1.55044763	0.8876	0.8876
2	0.22477619		0.1124	1.0000

Eigenvectors

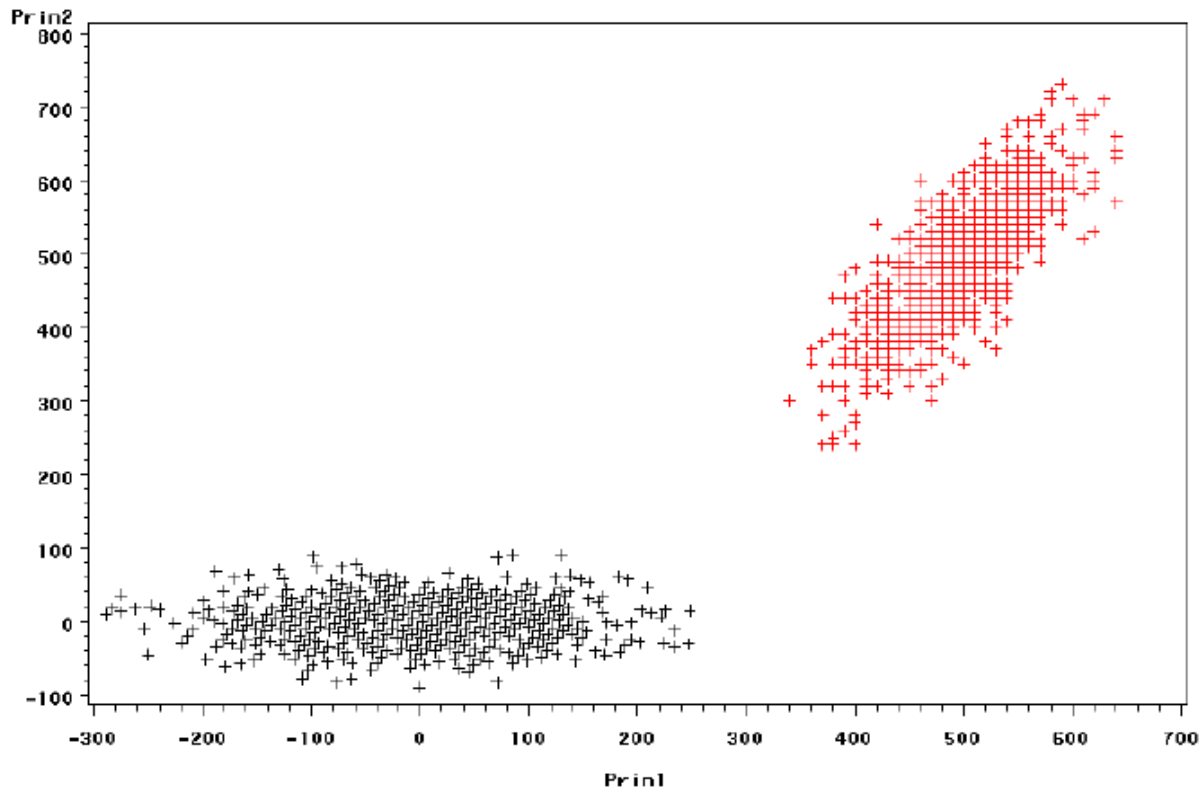|	Prin1	Prin2
SATV	0.707107	0.707107
SATM	0.707107	-.707107

Graphical Representation

- Plotting the components and the original data side by side reveals the nature of PCA:

  ➢ Shown from PCA of covariance matrix

PCA Summary

- This introduction to PCA was meant to help describe the nature of eigenvalues and eigenvectors
  - Many classical multivariate statistics use these terms

- We will return to PCA in the last week of class – when we discuss inferences about covariances in multivariate data using exploratory and confirmatory factor models

- Eigenvalues and eigenvectors will reappear in the next 2-3 weeks
  - The basis for classical Multivariate ANOVA
  - Modern versions have moved away from these terms
    - But they still are useful
  - They do show up from time to time in other places as well

WRAPPING UP

Wrapping Up

- Matrix algebra is the language of multivariate statistics
  - Learning the basics will help you read work (both new old)

- Over the course of the rest of the semester, we will use matrix algebra frequently
  - It provides for more concise formulae

- In practice, we will use matrix algebra very little
  - But understanding how it works is the key to understanding how statistical methods work and are related

- Up next: multivariate statistical distributions
  - The basis for modern estimation methods of multivariate statistical models using maximum likelihood or Bayesian estimation