Introduction to Mathematics for Political Science: Inner Product Spaces, Orthogonality, Projection

4 September 2018

Suppose we observed m features related to development across a sample of n countries – GDP per capita, life expectancy, literacy rate, etc. We're interested in creating some "index of development" that describes these data with a single number. These observations can be thought of as elements of a set. We started our study of linear spaces with a motivating question. How can we summarize many statistics related to a country's development with a single number, an "index of development"? We learned in the last lecture that a linear space with an inner product is an inner product space. Today, we put inner product spaces to work in order to construct our index of development.

Orthogonality

You studied orthogonality when learning about vectors and matrices. Two vectors are *orthogonal* if their inner product is zero. Geometrically, vectors are orthogonal when they are perpendicular, making a 90 degree angle with one another. The process of dimension reduction necessary to generate our Development Index is really about finding a set of orthogonal linear subspaces that best "fit" the data, and defining our data in terms of those subspaces. This will make more sense once we understand better orthogonality and projection.

Definition: Two *vectors* $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* iff $\boldsymbol{x}^T\boldsymbol{y} = 0$.

Theorem (Pythagorus): $\boldsymbol{x}^T\boldsymbol{y} = 0 \implies \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 = \|\boldsymbol{x}+\boldsymbol{y}\|^2$

Definition: Two *spaces* X and Y are orthogonal if for all $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$, $\boldsymbol{x}^T\boldsymbol{y} = 0$.

Definition: An *orthogonal basis*[1] for a linear space X is a set of basis vectors $S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ such that for all basis vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with $i \neq j$, $\boldsymbol{x}_i^T\boldsymbol{x}_j = 0$.

Remark: The standard basis is an orthogonal basis.

Example: Name a few orthogonal bases for $\mathbb{R}^2$.

Definition: An *orthonormal basis* is an orthogonal basis composed entirely of unit vectors.

Projection

Now we get to put all these concepts to use, and start thinking about how to move our data between different linear subspaces. The idea

[1] Remember, a *basis* for a linear space X is a set $\boldsymbol{y}_1, ..., \boldsymbol{y}_n \in S$ such that for all $\boldsymbol{x} \in X$,

$$\alpha_1\boldsymbol{y}_1 + ... + \alpha_n\boldsymbol{y}_n = \boldsymbol{x}$$

behind *projection* is to take an element of a linear space, and to push it into some linear subspace of that space.

Let S be a subspace of a normed linear space X and take some $\boldsymbol{x} \in X$. Let $\boldsymbol{p}_s : X \to S$ be a function that maps this $\boldsymbol{x} \in \mathbb{R}^n$ to its nearest point in S. Formally,[2]

$$d(\boldsymbol{x}, \boldsymbol{p}_s(\boldsymbol{x}) = d(\boldsymbol{x}, S) = \min \{d(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{y} \in S\}$$

How are orthogonality and projection related? If $\boldsymbol{x}$ and $\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x})$ are orthogonal, then $\boldsymbol{p}_s(\boldsymbol{x})$ must be a projection (and vice versa. To see why, take some other $\boldsymbol{y} \neq \boldsymbol{x} \in S$. We have

$$\|\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x})\|^2 \leq \|\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x})\|^2 + \|\boldsymbol{p}_s(\boldsymbol{x}) - \boldsymbol{y}\|^2$$

Since $\boldsymbol{p}_s(\boldsymbol{x}) - \boldsymbol{y} \in S$ and $\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x}) \perp S$, Pythagorus's Theorem applies

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x})\|^2 &\leq \|\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x})\|^2 + \|\boldsymbol{p}_s(\boldsymbol{x}) - \boldsymbol{y}\|^2 \\ &= \|\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x}) + \boldsymbol{p}_s(\boldsymbol{x}) - \boldsymbol{y}\|^2 \\ &= \|\boldsymbol{x} - \boldsymbol{y}\|^2 \\ \|\boldsymbol{x} - \boldsymbol{p}_s(\boldsymbol{x})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\| \end{aligned}$$

allowing us to conclude $d(\boldsymbol{x}, \boldsymbol{p}_s(\boldsymbol{x})) \leq d(\boldsymbol{x}, \boldsymbol{y})$ for all $\boldsymbol{y} \in S$.

This formal definition helps us see why orthogonality plays such a crucial role in projection. If $\boldsymbol{x} - P\boldsymbol{x}$ is orthogonal to S, we know that we're taking the shortest path from $\boldsymbol{x}$ to S, satisfying the requirements of a projection.[3]

Example: Consider the vector $\boldsymbol{a} = (1, 2, 3)$. This vector occupies $\mathbb{R}^3$. Suppose we wanted to *project* this vector onto the z axis, a one dimensional subspace of $\mathbb{R}^3$. To do so, we simply multiply by the matrix

$$P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which takes third component of $\boldsymbol{a}$ and drop the other two components, leaving $P\boldsymbol{a} = \boldsymbol{p}_3 = (0, 0, 3)$. Suppose we wanted to project into the xy-plane? Here, we simply let,

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

dropping the z component, and giving $P\boldsymbol{a} = \boldsymbol{p}_{12} = (1, 2)$ What's special about P and $\boldsymbol{p}$?

1. Projection matrices are *symmetric*: $P^T = P$
2. Projection matrices are *idempotent*: $PP = P$
3. Projections $\boldsymbol{p}$ are stable: $P\boldsymbol{p} = \boldsymbol{p}$[4]

[2] Can you prove that $\boldsymbol{p}_s$ is idempotent, given your knowledge of metric spaces?

[3] If S is a convex and compact set and X is a Euclidean Space, then the projection operator is a continuous function. Think about why this is the case. Can you prove it?

[4] Verify these properties for the example above.

For empirical political scientists this is not an academic conversation. Return to the cross national data discussed at the beginning of the lecture. We observe N countries and M features of those countries. Suppose we wanted to know how life expectancy varies with literacy rates and per capita gdp. We can store life expectancy observations in a vector $\boldsymbol{y}_{1\times N}$. We can store the other observations in a matrix $A_{N\times 2}$. In a simple linear model, we're interested in finding $\boldsymbol{\beta}_{1\times 2}$ that relates per capita gdp and literacy to life expectancy. So we could attempt to solve

$$\boldsymbol{y} = \boldsymbol{\beta} A^T$$

With $N > M - 1$, this system is unlikely to have a solution unless the data exhibit severe collinearity. We solve this problem through projection. We'll explore this in more detail in tomorrow's lecture. But you're now equipped with the tools to think about *how* to solve it. $\boldsymbol{\beta}$ defines a linear subspace of the N-dimensional space of rows. If we *project* $\boldsymbol{y}$ into this subspace, then we can take linear combinations of the columns of A and reach $\boldsymbol{y}$.[5]

Let's practice with a simple example.

Example: Now, suppose we wanted and orthogonal projection of an element in two dimensional space onto some line passing through the origin.[6] Call this line the vector $\boldsymbol{a}$ and the element with the vector $\boldsymbol{b}$. We know the projection will lie along this line

$$\boldsymbol{p} = \hat{x}\boldsymbol{a}$$

where $\hat{x}$ is some scalar. Let $\boldsymbol{e}$ represent the vector connecting $\boldsymbol{p}$ and $\boldsymbol{a}$

$$\boldsymbol{e} = \boldsymbol{b} - \hat{x}\boldsymbol{a}$$

We know from our definition of an orthogonal projection that this vector is orthogonal to $\boldsymbol{a}$, or

$$(\boldsymbol{a})^T(\boldsymbol{b} - \hat{x}\boldsymbol{a}) = 0$$

Now apply your linear algebra skills to solve for $\hat{x}$. We see

$$\hat{x} = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}$$

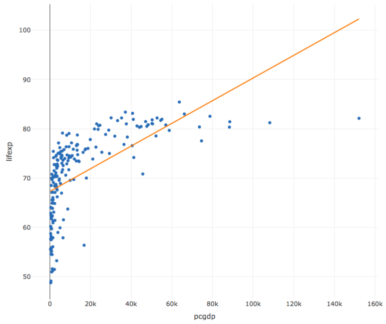
Our projection is

$$\boldsymbol{p} = \hat{x}\boldsymbol{a} = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}\boldsymbol{a}$$

What does this tell us about the projection matrix P?

Figure[7] demonstrates this algebra and each component of the projection visually.

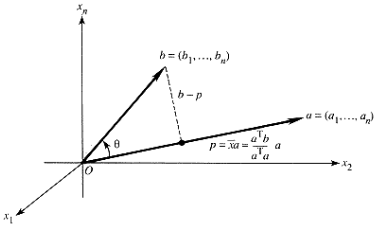
[5] Think about this discussion in the context of a simple 2-dimensional linear regression.



To where are the observations in $\boldsymbol{y}$ projected?

[6] For the more general case of projection onto a subspace, see Strang p. 209. This exercise should give you the intuition for this case.

[7]



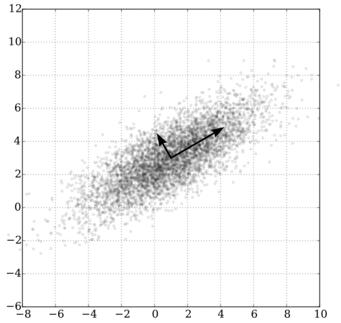
Application: Principal Component Analysis (PCA)

Principal component analysis (PCA) is a method used to "discover" underlying dimensions in data. Using the language we've learned today, this method iteratively finds one-dimensional *linear subspaces* that best explain the variance in the data, or minimize the distance between these subspaces and the elements of the data. We can think of this as choosing a "most explanatory" *orthogonal basis* for our data. The *projections* of these data into these linear subspaces are called component scores and are interpreted in relation to the underlying data.

We can think of the problem of developing an Development Index as a constrained prediction problem. If I asked you to predict the economic, health, and education outcomes of a country on the basis of a single number, one for each country, what number would do the best job? Arguably, this number would be a good candidate for our development index, and principal components analysis provides us with a tool to recover that number.

Implementing PCA requires an understanding of eigenvalues and matrix decompositions – material you'll be exposed to in later courses. But our work today should help you understand the intuition for the procedure. PCA selects an *orthogonnal basis*, rotating and scaling our standard coordinate system to better fit the data, and then projects these data into the orthogonal basis. You can see this in action in Figure[8].

Finally, I grabbed some data on our outcomes of interest from the World Bank. Life expectancy, GDP per capita, and literacy rates, and conducted principal components analysis on them. The component scores for the first dimension are visualized in Figure 4.



[8]

Exercises

1. Let X and Y be normed linear spaces. Let $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ be a basis for X and $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_m\}$ a basis for Y. Prove that if $x_i \perp y_j$ for all $i \in \{1, ..., n\}$, $j \in \{1, ..., m\}$, then X and Y are orthogonal spaces.

Because $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ is a basis for X, all $\boldsymbol{x} \in X$ can be expressed as linear combinations of $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. We therefore need to show that all linear combinations of $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ are orthogonal to all linear combinations of of $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_n\}$. More formally, we need

$$\alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j \perp \alpha_y \boldsymbol{y}_i + \beta_y \boldsymbol{y}_j$$

for arbitrary $\alpha_x, \beta_x, \alpha_y, \beta_y, \boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{y}_i, \boldsymbol{y}_j$. Equivalently,

$$\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i + \beta_y \boldsymbol{y}_j \rangle$$

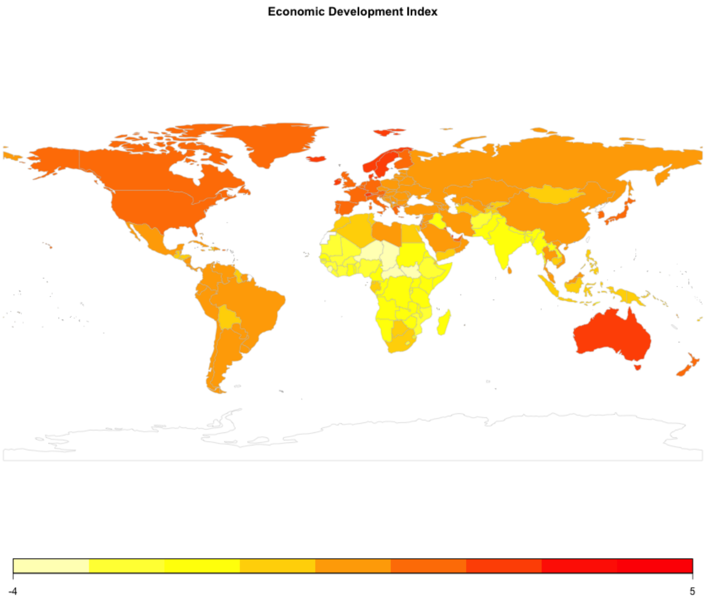


Figure 1: Development Index. Click here to view replication code.

By the bilinearity (additivity) of the inner product, we can write

$$\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i + \beta_y \boldsymbol{y}_j \rangle = \underbrace{\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i \rangle}_{A} + \underbrace{\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \beta_y \boldsymbol{y}_j}_{B}$$

Focusing on A, we also know by the symmetry of inner products that

$$\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i \rangle = \langle \alpha_x \boldsymbol{x}_i, \alpha_y \boldsymbol{y}_i \rangle + \langle \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i \rangle$$

Again by the bilinearity (homogeneity) of the inner product, this can be written

$$\langle \alpha_x \boldsymbol{x}_i, \alpha_y \boldsymbol{y}_i \rangle + \langle \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i \rangle = \alpha_x \alpha_y \underbrace{\langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle}_{=0} + \beta_x \alpha_y \underbrace{\langle \boldsymbol{x}_j, \boldsymbol{y}_i \rangle}_{=0}$$

where $\langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle$ and $\langle \boldsymbol{x}_j, \boldsymbol{y}_i \rangle = 0$ by the orthogonality of the basis vectors. We conclude

$$\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i \rangle = 0$$

Repeating the same argument for B gives

$$\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \beta_y \boldsymbol{y}_j \rangle = 0$$

which gives

$$\langle \alpha_x \boldsymbol{x}_i + \beta_x \boldsymbol{x}_j, \alpha_y \boldsymbol{y}_i + \beta_y \boldsymbol{y}_j \rangle = 0$$

as desired. ■

2. Prove: If a vector $\boldsymbol{\alpha}$ is in the null space of a set of vectors $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, then it is orthogonal to the space spanned by $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_m\}$ where

$$\boldsymbol{y}_i = \{x_{1i}, ..., x_{ni}\}$$

Let

$$X = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{bmatrix}$$

store the $\boldsymbol{x}$ vectors. If $\boldsymbol{\alpha} \in N(X)$, then

$$X\boldsymbol{\alpha} = \mathbf{0}$$

or

$$\boldsymbol{\alpha}^T X^T = \mathbf{0}$$

Now let

$$X^T = \begin{bmatrix} \boldsymbol{y}_1 & \cdots & \boldsymbol{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}$$

We can represent the space spanned by these vectors with $X^T\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ taking arbitrary linear combinations of the columns of X^T. We want to show

$$\boldsymbol{\alpha}^T X^T \boldsymbol{\beta} = \mathbf{0}$$

Because $\boldsymbol{\alpha}^T X^T = \mathbf{0}$, this must be the case. ■.

3. Donald Trump tweeted 100 times in April, 150 times in May, and 110 times in June.[9] Let $\boldsymbol{b} = (100, 150, 110)$ represent the number of tweets in each month. Project $\boldsymbol{b}$ onto $\boldsymbol{a} = (1, 1, 1)$. Interpret your result.

We have $\boldsymbol{p} = \hat{x}\boldsymbol{a}$ and $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$. We need $\boldsymbol{e}$ to be orthogonal to $\boldsymbol{a}$, or

$$\begin{aligned} \boldsymbol{a}^T(\boldsymbol{b} - \hat{x}\boldsymbol{a}) &= 0 \\ \boldsymbol{a}^T\boldsymbol{b} - \hat{x}\boldsymbol{a}^T\boldsymbol{a} &= 0 \\ \boldsymbol{a}^T\boldsymbol{b} &= \hat{x}\boldsymbol{a}^T\boldsymbol{a} \\ \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}} &= \hat{x} \end{aligned}$$

Substituting our values, this becomes

$$\frac{\sum_i b_i}{\sum_i 1} = \frac{100 + 150 + 110}{3} = 120$$

Notice that for n months of tweeting, this is

$$\frac{1}{n}\sum_i b_i$$

or simply the mean number of tweets.

[9] Disclaimer: these data are of suspect quality.

References

- Moore, Will H. and David A. Siegel. *A Mathematics Course for Political and Social Research*. Chapter 13.
- Strang, Gilbert. *Introduction to Linear Algebra*. Chapters 4, 6.
- Carter, Michael. *Foundations of Mathematical Economics*. Chapter 3.4, 3.6.
- Ok, Efe A. *Real Analysis with Economic Applications*. Chapter D.