Project Draft

Serena Cai, Alvin Yao, Cindy Wang

November 2020

Introduction

The COVID-19 pandemic has taken the United States by storm, appearing to worsen more and more every day with the end still very far from sight. Over the past few months, organizations, researchers, healthcare professionals, and our government have collected and made available large amounts of data on everything from the number of hospital beds per state to every policy response enacted in the state of New York.[8] Pinpointing which populations are most susceptible to the virus, which underlying factors cause the greatest variance, and how all of our actions and societal structures may affect each other is essential to help us end the pandemic and better prepare our global community to prevent future ones. However, with so many variables at hand, how can we identify the ones that truly matter? This is where PCA, or Principle Component Analysis, comes in.

With the growing power of technology to compute and store vast amounts of information, it feels as if there is so much to analyze and parse. However, having too many dimensions to data can be more detrimental than beneficial — the COVID-19 pandemic data being an example of that. High-dimensional data can slow down algorithms and be ultimately more distracting than intended.

Principal component analysis (PCA) is a key tool to reduce the dimensionality of data while maximizing its original variance. Applied as a preliminary step in a diverse array of fields from economics to the life sciences to study how a large number of variables affect target populations, PCA is primarily employed to remove unnecessary or distracting variables to help researchers focus on the variables that play the most important roles in a given issue. It can highlight the relative importance of each variable in determining the distribution and clustering of similar populations.[4]

In this paper, we will first explore the process of PCA through an example subset of US Census Data, diving into statistical context and linear algebra background as necessary. Then, we will discuss the different ways in which PCA achieves dimensionality reduction before diving into applications of PCA to the current COVID-19 pandemic. To conclude the paper, we hope to emphasize the importance and limitations of PCA and its power in serving as an excellent data analysis tool.

Statistical Context

Before we jump into PCA, let us cover a few key principles of statistical analysis. Generally speaking, we want to analyze how a single variable changes for a set of objects. Let us suppose that we are interested in seeing how the average income per household, represented by the single variable Q, varies over all of the countries in the world. However, this Q could represent any variable we would like to measure for a given target population. In an ideal world, we would have an average income per household for each of the countries in the world, the data for the entire population.

Accordingly, if we wanted to find the average income per household worldwide, we would take the sum of the average income per household for each country and divide it by the number of countries. This value, denoted by μ_Q, is our **population mean**

Population Mean

$$\mu_Q = \frac{Q_1 + Q_2 + ... + Q_n}{n}$$

The population mean is useful to calculate in that it helps us identify the center of the distribution of our data. This will become important in our PCA when we want to transform our data so that the center of the data is at the origin for ease of use for our following linear transformations and change-of-bases.

To center the data around the population mean, we can subtract each of the objects in the set we are analyzing by the population mean for that variable to find the **mean-deviation form**.[5]

Mean-Deviation Form

$$\begin{bmatrix} Q_1 - \mu_Q & \dots & Q_n - \mu_Q \end{bmatrix}$$

Suppose we also wanted to quantify how the data is spread around the mean, the **variance** of average income per household per country. To find variance, we calculate the average of the squared differences of each country's average income per household and the population mean. The differences are squared to ensure that all measurements are positive and measurements from one side of the mean cannot cancel out those from the other side.

Variance

$$Var = \frac{\Sigma(x - \mu)^2}{n}$$

In our scenario, since each country's data point for average income per household is represented by Q_i, for i such that $1 \leq i \leq n$, the **population variance** would be

Population Variance

$$Var(Q) = \frac{\Sigma(Q_i - \mu)^2}{n}$$

Interestingly, since the variance is calculated from squared differences, the variance is always in $units^2$. As the variance is often used in statistical calculations to determine how spread apart objects represented by datapoints are for a particular measurement, we can take the square root of the variance to find the **standard deviation** of the data.

Unfortunately, we often cannot obtain all of the data for a given population and must use a smaller subset of data, referred to as a **sample population**. In our example, this is analogous to not being able to collect the average income per household for all countries. As the number of data points in a sample population increases, the statistical calculations on the sample population increase in similarity to those of the population. Since the sample is an estimate of the population or actual intended target of study, there are small changes we have to make to calculate the variance of the sample population.

The **sample mean**, represented by $\bar{x}$, is calculated by taking the average of all of the data points in the sample.

Sample Mean
For a sample of size n,

$$\bar{x} = \frac{Q_1 + ... + Q_n}{n}$$

It turns out that the sample mean is always an underestimate of the population mean due to the way in which the sample mean is calculated as a subset of the population mean.[3] To account for this underestimation, the **sample variance** is calculated as follows:

Sample Variance

$$Var(Q) = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

What if we wanted to see not only how one variable varies among a population, but multiple variables and if there exists some relationship between such variables? Suppose we add another variable, P, representing the percent with health insurance coverage in a country for all countries in the world. We can then create a $2 \times n$ matrix containing the sample data (with n samples) for Q, P to be:

$$\begin{bmatrix} Q_1 & Q_2 & ... & Q_n \\ P_1 & P_2 & ... & P_n \end{bmatrix}$$

To capture and categorize the relationship between Q, P, we can calculate the **covariance** of Q, P, by evaluating the **sample covariance**, $Cov(Q, P)$.

Sample Covariance

$$Cov(Q, P) = \frac{\Sigma(Q_i - \bar{x_Q})(P_i - \bar{x_P})}{n - 1}$$

When the covariance is positive, this indicates that as Q and P are **positively correlated**. Thus, as Q increases, P increases. When the covariance is zero, Q and P have no correlation. And when the covariance is negative, Q and P are **negatively correlated**. Hence, when Q increases, P decreases, and vice versa.

Finally, we can represent the variances and covariances between all of the variables measured in a matrix through the **sample covariance matrix**, S. Continuing our example with Q, P, our covariance matrix would be a 2×2 matrix since we only have two variables, Q, P, such that:

$$\begin{bmatrix} Cov(Q,Q) & Cov(Q,P) \\ Cov(P,Q) & Cov(P,P) \end{bmatrix} = \begin{bmatrix} Var(Q) & Cov(Q,P) \\ Cov(P,Q) & Var(P) \end{bmatrix}$$

$$Cov(Q, P) = Cov(P, Q)$$

To describe the covariance matrix generally for any set of m variables and n samples, with our $m \times n$ matrix B containing the sample data in mean-deviation form with variables as rows,

$$B = \begin{bmatrix} A_1 - \bar{x_A} & A_2 - \bar{x_A} & \dots & A_n - \bar{x_A} \\ \vdots & \vdots & \ddots & \\ M_1 - \bar{x_M} & M_2 - \bar{x_M} & \dots & M_n - \bar{x_M} \end{bmatrix}$$

our covariance matrix S will be defined as:

Covariance Matrix

$$S = \frac{1}{n-1} BB^T$$

Since the covariance matrix should represent the relationships between all of the variables, its number of rows and columns should equal the number of variables. Thus, taking a starting matrix with different samples as columns and distinct variables as rows, we can obtain a matrix with such dimensions by multiplying our starting matrix by its transpose. If we expand our definition of the covariance matrix with an example matrix B, we can see why we need to multiply B by its transpose.

$$S = \frac{1}{n-1} \begin{bmatrix} A_1 - \bar{x_A} & A_2 - \bar{x_A} & \dots & A_n - \bar{x_A} \\ \vdots & \vdots & \ddots & \\ M_1 - \bar{x_M} & M_2 - \bar{x_M} & \dots & M_n - \bar{x_M} \end{bmatrix} \begin{bmatrix} A_1 - \bar{x_A} & \dots & M_1 - \bar{x_M} \\ A_2 - \bar{x_A} & \dots & M_2 - \bar{x_M} \\ \vdots & \ddots & \vdots \\ A_n - \bar{x_A} & \dots & M_n - \bar{x_M} \end{bmatrix}$$

$$= \begin{bmatrix} Var(A) & \dots & Cov(A,M) \\ \vdots & \ddots & \vdots \\ Cov(M,A) & \dots & Var(M) \end{bmatrix}$$

We will further explore how the variances and covariances are represented in the covariance matrix in the following example.

Example

Suppose we have three states, x_1, x_2, x_3, representing Arkansas, California, and Colorado, respectively. Using data collected from the US Census Bureau[1], we are measuring four variables—percent of population with no health insurance coverage, percent paid below poverty level, percent of households where grandparents are primary caregivers, percent female population—on which we measure three characteristics. Our datapoints representing each of the three states, our **observation vectors** are as follows:

$$x_1 = \begin{bmatrix} 9.0 \\ 17.6 \\ 0.0127 \\ 0.504 \end{bmatrix}, x_2 = \begin{bmatrix} 8.5 \\ 14.3 \\ 0.00651 \\ 0.498 \end{bmatrix}, x_3 = \begin{bmatrix} 8.1 \\ 10.9 \\ 0.00605 \\ 0.478 \end{bmatrix},$$

Problem: Find the sample mean, mean-deviation form, and construct the B matrix and the sample covariance matrix.

Solution: First, to find the sample mean:

$$\mathbf{M} = \frac{1}{3}\left(\begin{bmatrix} 9.0 \\ 17.6 \\ 0.0127 \\ 0.504 \end{bmatrix} + \begin{bmatrix} 8.5 \\ 14.3 \\ 0.00651 \\ 0.498 \end{bmatrix} + \begin{bmatrix} 8.1 \\ 10.9 \\ 0.00605 \\ 0.478 \end{bmatrix}\right) = \frac{1}{3}\left(\begin{bmatrix} 25.6 \\ 42.8 \\ 0.02526 \\ 1.48 \end{bmatrix}\right) = \begin{bmatrix} 8.53 \\ 14.27 \\ 0.00842 \\ 0.493 \end{bmatrix}$$

Then, to find the mean-deviation form, let us subtract the sample mean from each observation vector.

$$\hat{x_1} = \begin{bmatrix} 0.47 \\ 3.33 \\ 0.00428 \\ 0.011 \end{bmatrix}, \hat{x_2} = \begin{bmatrix} -0.03 \\ 0.03 \\ -0.00191 \\ 0.005 \end{bmatrix}, \hat{x_3} = \begin{bmatrix} -0.43 \\ -3.37 \\ -0.00237 \\ -0.014 \end{bmatrix}$$

Our B matrix would then be as follows:

$$\begin{bmatrix} 0.47 & -0.03 & -0.43 \\ 3.33 & 0.03 & -3.37 \\ 0.00428 & -0.00191 & -0.00237 \\ 0.011 & 0.005 & .-0.014 \end{bmatrix}$$

To construct our sample covariance matrix,

$$S = \frac{1}{2} \begin{bmatrix} 0.47 & -0.03 & -0.43 \\ 3.33 & 0.03 & -3.37 \\ 0.00428 & -0.00191 & -0.00237 \\ 0.011 & 0.005 & -0.014 \end{bmatrix} \begin{bmatrix} 0.47 & 3.33 & 0.00428 & 0.011 \\ -0.03 & 0.03 & -0.00191 & 0.005 \\ -0.43 & -3.37 & -0.00237 & -0.014 \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} 0.4067 & 3.0133 & 0.003088 & 0.01104 \\ 3.0133 & 22.4467 & 0.022182 & 0.08396 \\ 0.003088 & 0.022182 & 0.0000275834 & 0.00007071 \\ 0.01104 & 0.08396 & 0.00007071 & 0.000342 \end{bmatrix} = \begin{bmatrix} 0.20335 & 1.50665 & 0.001544 & 0.00552 \\ 1.50665 & 11.22335 & 0.011091 & 0.04198 \\ 0.001544 & 0.011091 & 0.0000137917 & 0.000035355 \\ 0.00552 & 0.04198 & 0.000035355 & 0.000171 \end{bmatrix}$$

From our sample covariance matrix, we can see that the entry $S_{1,1}$ is

$$S_{1,1} = \frac{1}{2}((9.0\text{–}8.53)^2 + (8.5\text{–}8.53)^2 + (8.1\text{–}8.53)^2 = 0.20335$$

And this is exactly the variance of what we declared as the first variable we measured on our observation vectors. And then, if we look at the $S_{2,1}$ entry

$$S_{2,1} = \frac{1}{2}((9.0-8.53)(17.6-14.27)+(8.5-8.53)(14.3-14.27)+(8.1-8.53)(10.9-14.27)) = 1.50665$$

We see that this is precisely the covariance of the first and second variables.

Following these patterns, we can generalize what we have observed to find that for $1 \leq i, j \leq m$:

1. On the diagonal of S, the ith entry, $S_{i,i}$ represents the variance of the ith variable.
2. Within S, $S_{i,j}$, where $i \neq j$, represents the covariance between the ith and jth variables.

Thus, our covariance matrix neatly packages all of the variances and covariances between our various measurements or variables for the given observation vectors.

Note: Our covariance matrix is symmetric.

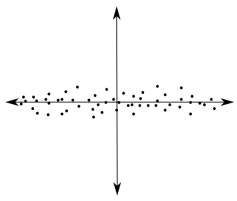


Figure 1: 2D plot with no correlation[31]

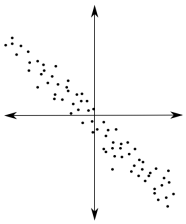


Figure 2: 2D plot with negative correlation[32]

Example 1: The covariance matrix can also be interpreted visually. Suppose we have a 2D plot as seen on the left.

Since the data points are scattered all along the x-axis, we can expect S_{11} to be very large. On the other hand, there is a tighter constraint on the y-axis. We can expect S_{22} to be smaller. We do not know much about the covariance because there is no correlation between the two variables. In other words, the position of the datapoint along its x variable component does not have much say in the position of that datapoint along its y component.

Let's look at another example where the data points appear to be diagonal. In this graph, the variance in both directions is essentially equal, $S_{11} = S_{22}$, and there is a strong negative correlation between the two variables. As the values along the horizontal axis increase, the values for the datapoints along the vertical axis increase. Thus, $S_{12} = S_{21} < 0$.

Finding Principal Components

Principal components are new variables constructed from linear combinations of the initial variables in ways that maximize the variance of the data. These combinations compress all the initial data into a few components on a lower dimension without loss of data. Hence, our principal components should be vectors that represent the directions where the data has the greatest variance.

In PCA, we derive our principal components from the covariance matrix of the variables and sample data. If our covariance matrix is $m \times m$ dimensional, suppose we multiply the covariance matrix with a random vector $v \in \mathbb{R}^m$, basically applying a matrix transformation with our covariance matrix on some vector. Every time we multiply our vector v by the covariance matrix, v will be transformed (notably not only scaling but turning/rotating) closer and closer towards a particular direction. This direction happens to be the direction of greatest variance. The direction of greatest variance, itself, will not be rotated or knocked off its span, but only scaled. By definition, eigenvectors are vectors that are only scaled by a matrix transformation or upon multiplication with a matrix.[5] Hence, the directions of greatest variance we seek to find are the eigenvectors of the covariance matrix.

However, we do not want to find any set of eigenvectors. We aim to find an **orthogonal** set of eigenvectors. The main idea of PCA is to first find the direction of greatest variance as the first principal component, before finding the next greatest direction of variance that is independent of the first.

Let's revisit the orthogonal decomposition theorem we proved in class[5].

The Orthogonal Decomposition Theorem
Let W be a subspace of $\mathbb{R}^n$. Then each y in $\mathbb{R}^n$ can be written uniquely in the form

$$y = \hat{y} + z$$

where $\hat{y}$ is in W and z is in $W^{\perp}$. In fact, if $\{u_1, ..., u_p\}$ is any orthogonal basis of W, then

$$\hat{y} = \frac{y \cdot u_1}{u_1 \cdot u_1} u_1 + \ldots + \frac{y \cdot u_p}{u_p \cdot u_p} u_p$$

and $z = y - \hat{y}$.

By the orthogonal decomposition theorem, we see that any vector space, such as $\mathbb{R}^n$, can be decomposed into a subspace and its orthogonal complement whose bases combine to span the entire vector space. This means that to find a principal component or eigenvector that is not "affected" by the earlier principal components/eigenvectors, we want to find eigenvectors that are not included in the span of the already-identified eigenvectors. Hence, we want to find eigenvectors that are orthogonal to each other.

By definition of an orthogonal set[5], we thus want to find an orthogonal diagonalization of the covariance matrix to find an orthogonal set of eigenvectors to be our principal components. By diagonalizing the covariance matrix, we can single out the eigenvalues on the diagonal that correspond to the largest variances in a particular direction. From these, we can obtain eigenvectors that are most representative of the data while knocking out the smaller variances that do not impact the data as much.

Orthogonal Diagonalizability of the Covariance Matrix

To understand why we can orthogonally diagonalize any given covariance matrix, let us explore a few topics in linear algebra.

Symmetric Matrices

By definition, a symmetric matrix is one that is equal to its transpose.[5] In other words,

$$A = A^T$$

where A is some square matrix.

From this, it necessarily follows that A's main diagonal entries can have any real value, but every other entry must have a corresponding entry on the opposite side of the main diagonal that it is equal to.

Spectral Theorem

The Spectral Theorem for Symmetric Matrices[5]
An $n \times n$ symmetric matrix A has the following properties:
a. A has n real eigenvalues, counting multiplicities
b. The dimension of the eigenspace for each eigenvalue λ equals the multiplicity of λ as a root of the characteristic equation.
c. The eigenspaces are mutually orthogonal, in the sense that eigenvectors corresponding to different eigenvaules are orthogonal.
d. A is orthogonally diagonalizable.

Proposition 1: As proved in class, if A is symmetric, then any two of its eigenvectors that correspond to distinct eigenvalues (i.e. from different eigenspaces) are orthogonal.

Definition: A square nxn matrix A is **orthogonally diagonalizable** if there exists an orthogonal matrix P (where $P^{-1} = P^T =$) and a diagonal matrix D such that

$$A = PDP^T = PDP^{-1}$$

This diagonalization requires n linearly independent and orthonormal eigenvectors. Thus, if we take matrix A from above and look at its transpose:

$$A^T = (PDP^T)^T = P^{TT}D^T P^T = PDP^T = A$$

Since we have proven $A = A^T$, we know that to be orthogonally diagonalizable, A must be symmetric. Conversely, it follows that all symmetric matrices are orthogonally diagonalizable, as proved in class.

Furthermore, any symmetric matrix A must have n real eigenvalues. In other words, there exist real numbers (namely, the eigenvalues) $\lambda_1, ..., \lambda_n$ and orthogonal, non-zero real vectors (namely, the eigenvectors) $\vec{v_1}, ..., \vec{v_n}$ such that for each eigenvector, we have:

$$A\vec{v} = \lambda\vec{v}$$

These two properties are part of the **Spectral Theorem of Symmetric Matrices.**

Going back to the original example, we can orthogonally diagonalize our covariance matrix. We first calculate the eigenvalues and their corresponding eigenvectors. For very small numbers, it has been denoted as ± 0.000 in this case.

$$\lambda_1 \approx 11.4258, E_{\lambda 1} \approx \begin{bmatrix} -0.13306 \\ -0.991101 \\ -0.000980053 \\ -0.00370579 \end{bmatrix}, \lambda_2 \approx 0.00108935, E_{\lambda 2} \approx \begin{bmatrix} 0.98425 \\ -0.131792 \\ 0.0503993 \\ -0.106498 \end{bmatrix},$$

$$\lambda_3 \approx +0.000, E_{\lambda 3} \approx \begin{bmatrix} 0.112023 \\ -0.0185375 \\ -0.166435 \\ 0.979493 \end{bmatrix}, \lambda_4 \approx +0.000, E_{\lambda 4} \approx \begin{bmatrix} 0.0315725 \\ -0.00262564 \\ -0.984763 \\ -0.170991 \end{bmatrix}$$

The diagonal matrix would be:

$$S =$$

$$\begin{bmatrix} -0.13306 & 0.98425 & 0.112023 & 0.0315725 \\ -0.991101 & -0.131792 & -0.0185375 & -0.00262564 \\ -0.000980053 & 0.0503993 & -0.0185375 & -0.984763 \\ -0.00370579 & -0.106498 & 0.979493 & -0.170991 \end{bmatrix} \begin{bmatrix} 11.4258 & 0 & 0 & 0 \\ 0 & 0.000108935 & 0 & 0 \\ 0 & 0 & +0.000 & 0 \\ 0 & 0 & 0 & +0.000 \end{bmatrix}$$

$$\begin{bmatrix} -0.13306 & 0.98425 & 0.112023 & 0.0315725 \\ -0.991101 & -0.131792 & -0.0185375 & -0.00262564 \\ -0.000980053 & 0.0503993 & -0.0185375 & -0.984763 \\ -0.00370579 & -0.106498 & 0.979493 & -0.170991 \end{bmatrix}^{-1}$$

The diagonal values are the principal components.

Principle 1

Remark: If A is any mxn matrix, then it follows that AA^T and A^TA are symmetric and mxm and nxn respectively.

Thus, proposition 2 follows:

Proposition 2: The matrices AA^T and A^TA share the same nonzero eigenvalues.

Proof: Suppose $\vec{v}$ is a nonzero eigenvector of A^TA (or $\lambda \neq 0$). By definition,

$$(A^TA)\vec{v} = \lambda\vec{v}.$$

If we multiply both sides by A, we get:

$$A(A^TA)\vec{v} = A(\lambda\vec{v}).$$

By the associative property of matrix multiplication and scalar multiplication, we can rewrite this as:

$$AA^T(A\vec{v}) = \lambda(A\vec{v}).$$

By definition, this must mean that the vector represented by $A\vec{v}$ is an eigenvector of AA^T with eigenvalue λ. Since eigenvectors by definition cannot be zero, we must check that $A\vec{v}$ is nonzero. However, since we previously defined $\vec{v}$ to be nonzero and $\lambda \neq 0$, and from the original equation, we can conclude that $A\vec{v}$ cannot be 0.[3] Thus, we have proved that any nonzero eigenvalue λ of A^TA is also an eigenvalue of AA^T.

This proposition is very powerful in situations where the number of rows is drastically different from the number of columns, in which case we can take advantage of this idea. For example, suppose we have a 600 x 3 matrix A. If we want to find the eigenvalues of AA^T, whose dimensions are a whopping 600 x 600, we can simply look at the eigenvalues of 3 x 3 matrix A^TA. Since A^TA is symmetric, we know that it must have 3 real eigenvalues. From this proposition, these 3 eigenvalues will also belong to AA^T, and the remaining 497 eigenvalues of AA^T (AA^T is also symmetric and will thus have 600 eigenvalues) will be zero![3,5]

By this principle, going back to our example, instead of calculating the eigenvectors and eigenvalues of the original $\frac{1}{2}BB^T$ matrix, we can find the eigenvalues and eigenvectors of $\frac{1}{2}B^TB$ which will be the same nonzero eigenvalues as the original matrix.

$$R = \frac{1}{2}B^TB = \frac{1}{2}\begin{bmatrix} 11.3099393184 & 0.0858468252 & -11.4243641436 \\ 0.0858468252 & 0.0018286481 & -0.0882654733 \\ -11.4243641436 & -0.0882654733 & 11.5420016169 \end{bmatrix}$$

We then order the eigenvalues and their corresponding eigenvectors in non-increasing order. The next step of PCA involves getting the total variance of the data set.

Remark: The trace of a matrix is equal to the sum of its eigenvalues.

The trace of S is the sum of the diagonal entries of S, otherwise known as the total variance, or T, of the data.[5] This is the sum of the variances of all m variables, or $T = \lambda_1 + ... + \lambda_m$.

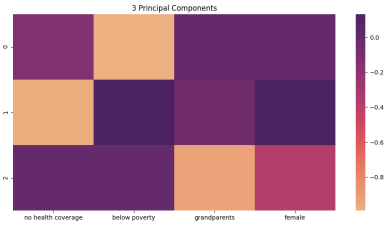
In order to determine how strongly a principle component accounts for variation within the data, we can simply take its corresponding eigenvalue and divide it by the trace (i.e. total variance). Thus, in our example, we can obtain the total variance by adding up the eigenvalues of the 4 eigenvectors as follows:

$$T = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 11.426893$$

Then, to measure how much each principle component influences the total variance, we can divide each corresponding eigenvalue by the total variance:

$$\begin{aligned}
\frac{\lambda_1}{T} &= \frac{11.4258}{11.426893} \approx 0.9999 \\
\frac{\lambda_2}{T} &= \frac{0.000108935}{11.426893} \approx 0.0001 \\
\frac{\lambda_3}{T} &= \frac{+0.000}{11.426893} \approx +0.000 \\
\frac{\lambda_4}{T} &= \frac{+0.000}{11.426893} \approx +0.000
\end{aligned}$$

Thus, the first principal component accounts for 99.99% of the total variance, while the second principal component accounts for 0.01%. The third and fourth principle components account for so little of the total variance that we can say with confidence that they have little to no impact on the total variance.



We also ran through this subset example in our Google Colab, and we created this heatmap that shows the weights applied to each variable for the first three principal components.[6−7,9] From this heatmap, we can see that the first principal component is largely determined by percent earning below the poverty level. This means that the greatest variance among states is in this factor, followed by percent without health coverage.

Dimension Reduction Possibilities

After calculating the eigenvalues and eigenvectors from a given data set, it is typical that the largest eigenvalues are much larger than the rest. In the example above, the first principal components would explain 99% of the total variation in the data. PCA would effectively reduce a data set in $\mathbb{R}_4$ down to $\mathbb{R}_1$ represented by the most significant feature.

The fundamental principles of PCA are as follows:

- The direction of $\vec{u}_1$ in $\mathbb{R}_m$ describes a fraction of the total variance, T. It can be written as $\frac{\lambda_1}{T}$. The second principal component $\vec{u}_2$ accounts for $\frac{\lambda_2}{T}$ of the total variance.
- The first principal component points in the most significant direction of the data set.
- Next, the second principal component would point in the most significant direction among directions that are orthogonal to $\vec{u}_1$.
- The third principal component would point in the most significant direction among directions that are orthogonal to both $\vec{u}_1$ and $\vec{u}_2$, and so forth.

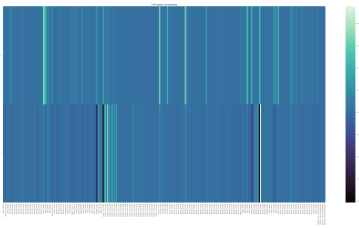
Application to the COVID-19 Pandemic

Now let us move on to applying PCA to examine how various demographic factors across the states of the United States have affected the cumulative cases and deaths per capita due to COVID-19.

We imported data from the New York Times COVID-19 Github repository on the number of COVID-19 cases and deaths over the course of the pandemic.[10] We also extracted data from the U.S. Census Bureau on demographics, housing, employment, and healthcare coverage of populations residing in each state from 2019 (pre-pandemic).[1] With these datasets, we created a 51×291 matrix with 51 observation vectors representing the 50 states and Puerto Rico and 291 demographic factors/preconditions.

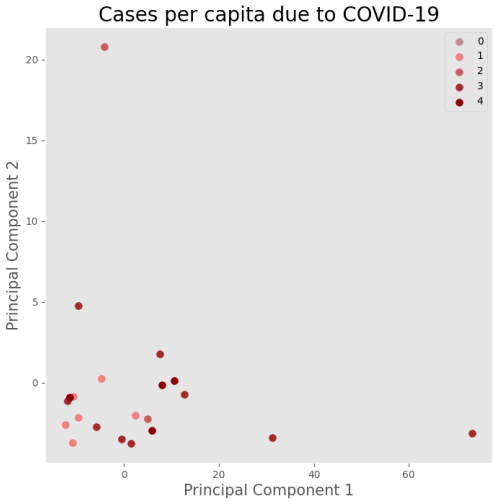
In this application, we used both the scikit-learn and numpy approaches to performing PCA on our dataset. A more specific and thorough run-down and tutorial of how we imported, cleaned, and applied PCA to our data can be found in our Google Colab. It is highly recommended to read through the notebook to fully understand our process.

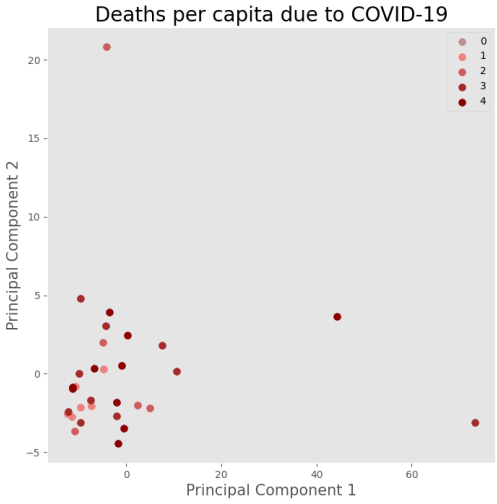
We first compiled all of our data from the US Census Bureau on our various measured variables into one pandas dataframe, on which we used the PCA package from scikit-learn to determine the principal components. The composition of the first two principal components can be seen in the following heatmap, where the vertical axis represents the principal components and the horizontal axis represents the various initial variables.



When examining the explained variance of the first two principal components, we found that just the first principal component accounts for more than 98 percent of the total variance. If we take a closer glance at some of the initial variables that had large weights in the linear combination forming the first principal component, we see that they include the population with health coverage, for whom poverty status is determined, and total state population.

After calculating our two principal component vectors and seeing that they account for 99.5 percent of the distribution of the data, we can project our data onto these two principal components as a scatter plot. In order to represent the varying levels of cases per capita due to COVID-19 and deaths per capita due to COVID-19, we found the quintiles for cases per capita and deaths per capita among the 51 regions and labeled





each state $\{0, 1, 2, 3, 4\}$, where 0 represents states with the lowest cases or deaths per capita and 4 represents the states with the highest cases or deaths per capita. Both graphs can be seen below.

From the distribution of colors on the graph, we can see that for cases per capita due to COVID-19, there appears to be clustering of states with higher levels of cases per capita when the first principal component is positive, while there is no such correlation in the deaths per capita graph. This may suggest that states with greater amounts of principal component 1 (thus, greater amounts of its component initial variables including healthcare coverage, poverty status, state population) experience greater cases per capita of COVID-19. One interpretation for this finding may be that areas with a greater wealth gap, such as largely populated states with urban areas, have greater healthcare coverage and a greater population of impoverished people and suffer greater cases per capita due to the disparity in quality of life and other structural inequalities.

By identifying the principal components and the contributions of each of the initial variables to the principal components, we can represent the data in a lower-dimensional format that is easier to analyze and remove any variables that may have just distracted from the real determinants of the distribution of the data.[4] In the context of the COVID-19 pandemic, this type of analysis can help public officials better target the underlying problems behind the spread and continued reign of the virus. By identifying what key preconditions and demographic factors contribute to different regions handling the pandemic differently, researchers can more effectively predict how different countries and states will respond to future pandemics.

Conclusion

In conclusion, PCA is a powerful dimension reduction tool that enables us to interpret data more easily by getting rid of redundancies and irrelevant variables while minimizing information loss. Of course, because of this feature, PCA also has its limitations such as initial independent variables becoming less interpretable and the necessity for the data to be standardized before PCA. Thus, PCA is only a preliminary step for most analyses, used in combination with other approaches like singular value decomposition.[4]

Despite these limitations, PCA has proven extremely useful regardless of the countless times it has been applied in fields that often entail large data sets, such as machine learning, finance, and bioinformatics. Because these fields tend to have data containing a lot of features that are often too many to effectively interpret and visualize, PCA can come in handy by compressing this much information to fewer dimensions while still retaining the most important factors.

In machine learning, for example, it is key that the deep learning program can accurately generalize inputs beyond what was used in the initial training of the program. However, as the dimensionality of the data increases, the ability of the program to decide what datapoints should be generalized decreases. Using PCA to decrease the dimensionality of training data helps preserve the ability of the program to accurately generalize data based on a maximal amount of variance. In finance, the clustering capabilities of PCA are helpful in putting similar stocks into the same principal component. These clusters allow users to pick one stock from each principal component and thus better diversify their investments to lower their overall risk. Some bioinformatics applications of PCA include large epidemiological studies such as our COVID-19 application and large gene expression pattern analyses.[4,7]

As our world becomes increasingly data-driven, despite its limitations, PCA is a key dimensionality-reduction technique that we all should embrace due to its wide applicability and efficacy in many important fields of study.

References

1 "Census COVID-19 Data Hub." Covid19.Census.Gov, US Census Bureau, 2020, covid19.census.gov/. Accessed 14 Dec. 2020. Datasets Used: ACS Labor Force Participation by Age - State; Computer and Internet Use - State; County Business Patterns - State; Disability Status of the Civilian Non-institutionalized Population - State; Grandparents Responsible for Grandchildren - State; Health Insurance Coverage - State; Households by Type - State; Housing Tenure - State; Income and Benefits - State; Language Spoken at Home - State; Non-Employer Statistics - State; Population and Poverty - State; Population by Age and Sex - State; Race and Ethnicity - State; School Enrollment - State; Worked at Home - State.

2 Galarnyk, M. (2020, October 17). PCA using Python (scikit-learn). Retrieved December 14, 2020, from https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

3 Jauregui, Jeff. (2012, August 31). Principal Component Analysis with Linear Algebra. Retrieved November 19, 2020 from http://www.math.union.edu/ jaureguj/PCA.pdf

4 Jolliffe, I. T, and Jorge C. "Principal component analysis: a review and recent developments." Philosophical transactions. Series A, Mathematical, physical, and engineering sciences vol. 374,2065 (2016): 20150202. doi:10.1098/rsta.2015.0202

5 Lay, D. Linear Algebra and its applications, 5th ed., Pearson, 2016

6 Navarrete, W. Principal Component Analysis with NumPy – Wendy Navarrete. 6 July 2020, wendynavarrete.com/principal-component-analysis-with-numpy/. Accessed 15 Dec. 2020.

7 Pramoditha, R. "Principal Component Analysis (PCA) with Scikit-Learn." Medium, 26 Oct. 2020, towardsdatascience.com/principal-component-analysis-pca-with-scikit-learn-1e84a0c731b0. Accessed 15 Dec. 2020.

8 Roser, M., Ritchie, H., Ortiz-Ospina, E., Hasell, J. (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus'

9 Sharma, A. "Principal Component Analysis (PCA) in Python." datacamp. Datacamp, January 1, 2020. https://www.datacamp.com/community/tutorials/principal-component-analysis-in-python.

10 Smith, M., et al. "Nytimes/Covid-19-Data." GitHub, New York Times, 4 Apr. 2020, github.com/nytimes/covid-19-data. Accessed 14 Dec. 2020.