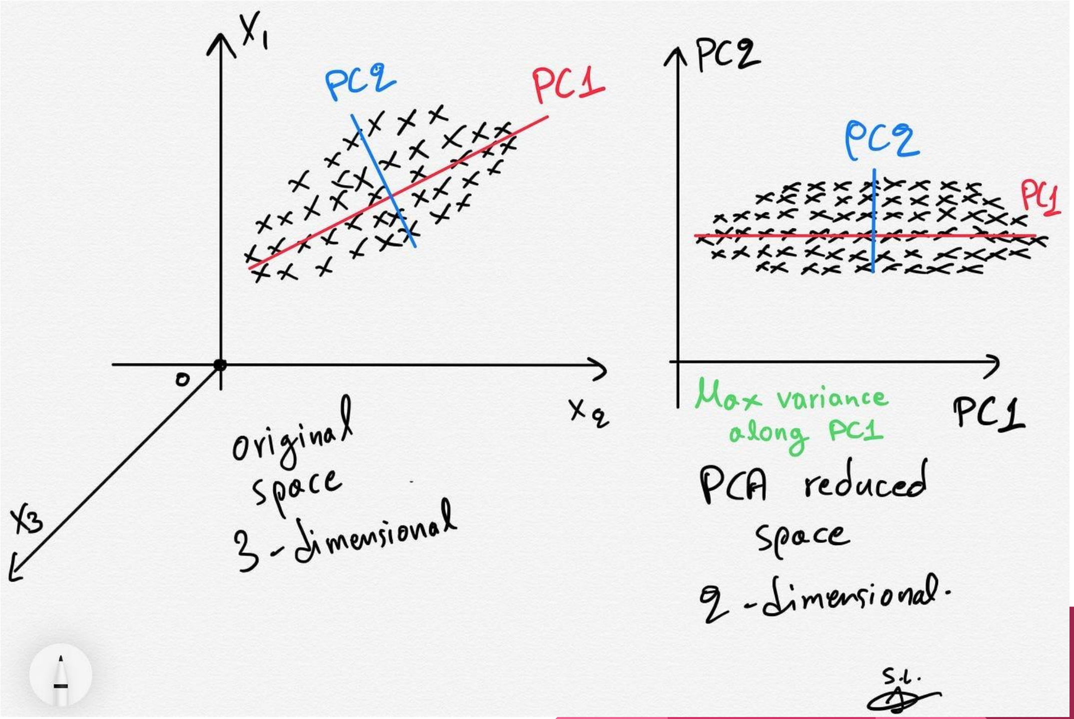
Principal Component Analysis with Spotify Data

Why use PCA?

- Reduces the dimensionality of data (helping performance of ML algorithms)
- Visualize data in 2 or 3 dimensions
- Helps find the main axes of variance
- Help find underlying features in data



Data Cleansing

- Converting columns to desired data types
- When working with data its important to scan through the data, weeding out any data that doesn't belong, such as:

```
[ ]  for
       if                                                   x)
```

```
track_name              object
artist(s)_name          object
artist_count             int64
released_year            int64
released_month           int64
released_day             int64
in_spotify_playlists     int64
in_spotify_charts        int64
streams                 object
in_apple_playlists       int64
in_apple_charts          int64
in_deezer_playlists     object
in_deezer_charts         int64
```

```
BPM1                lence75Energy69Acousticness7Instrumentalness0Liveness17Speechiness3
```

Standardizing the Data

- Before running PCA we standardize the data to have a mean of 0 and a standard deviation of 1
- This procedure subtracts the data by the mean and divides it by the standard deviation
- Purpose of this is to ensure the results are not biased by the scale of the data

$$Z = \frac{x - \mu}{\sigma}$$

Computing the Covariance Matrix

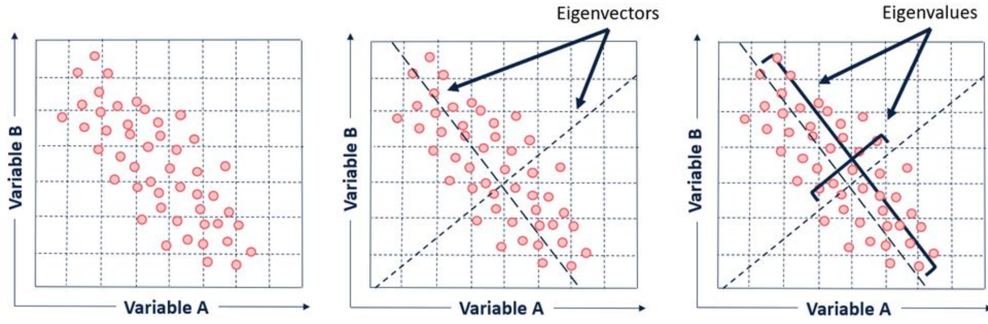
- The covariance matrix finds the correlation between features in the data

$$covar(f1, f2) = \frac{\sum_{i=1}^{n}(f1_i - \overline{f1})(f2_i - \overline{f2})}{n-1}$$

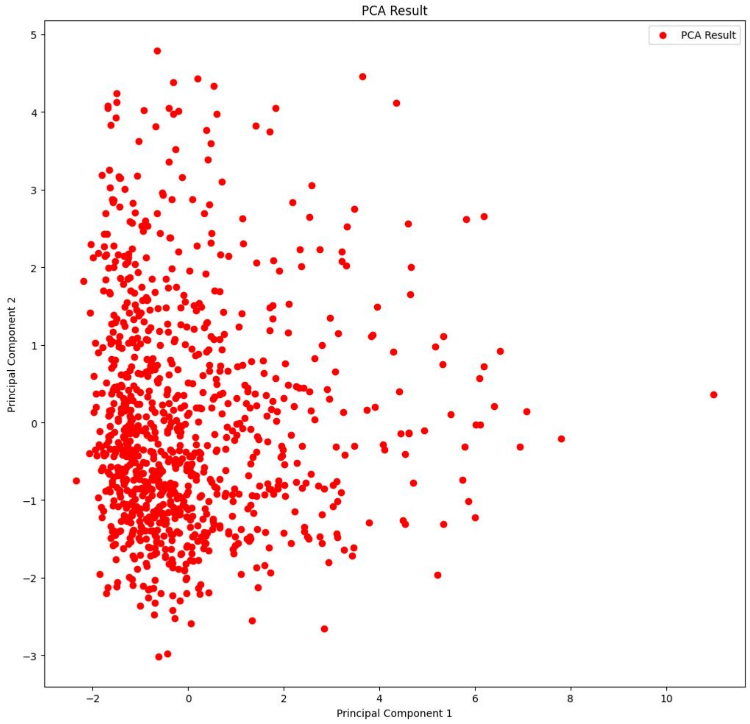
	Streams (f1)	BPM(f2)	...	(f_N)
Streams (f1)	covar(f1, f1)	covar(f1, f2)	...	covar(f1, f_N)
BPM (f2)	covar(f2, f1)	covar(f2, f2)	...	covar(f_N, f2)
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
(f_N)	covar(f_N, f1)	covar(f_N, f2)	...	covar(f_N, f_N)

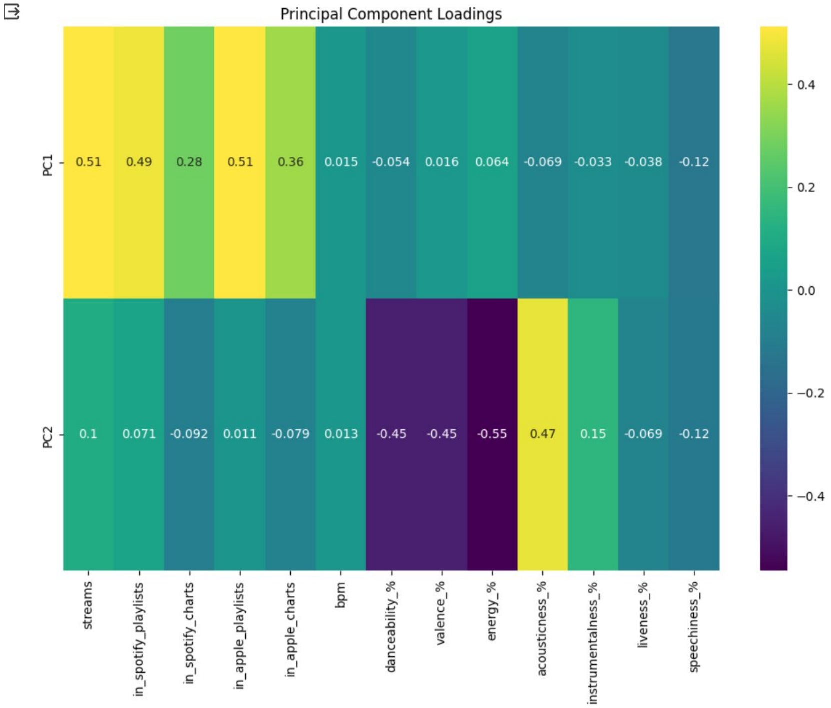
Eigenvectors and Eigenvalues to Find Principal Components

- Now we find the principal components that fit the highest variance of the data
- To find the principal components is equivalent to finding the eigenvalues and eigenvectors of the covariance matrix

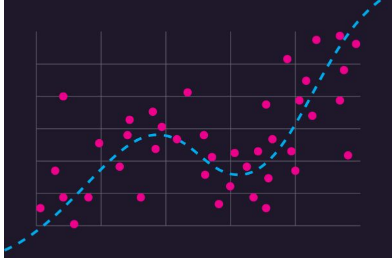


Recasting Data on Principal Components





Combining PCA and Regression



Principal component regression is a regression technique that has the same goal as standard linear regression which is to model the relationship between a target variable and the predictor variables

Our goal: Model the relationship between streams and the other integer variables (predictor variables)

Principal Component Regression Steps

1. Apply PCA to generate principal components from the predictor variables

2. Keep the first k principal components that explain most of the variance (where $k < p$), where k is determined by cross-validation

3. Fit a linear regression model on these k principal components

Pros/Cons

Pros:

Fits a linear regression model on k principal components instead of all the original features, thus helping to reduce overfitting

Eliminate multicollinearity in the data by removing principal components associated with small eigenvalues

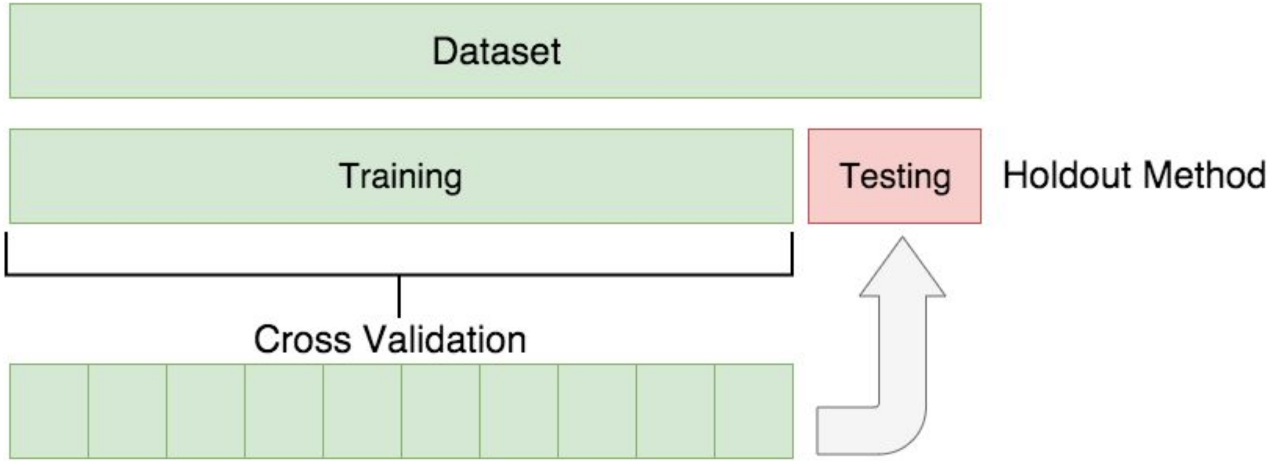
Cons:

Does not consider the target variable when determining principal components

Not considered a feature selection method because the principal components used in the regression are linear combinations of the original features

Train/Test Split

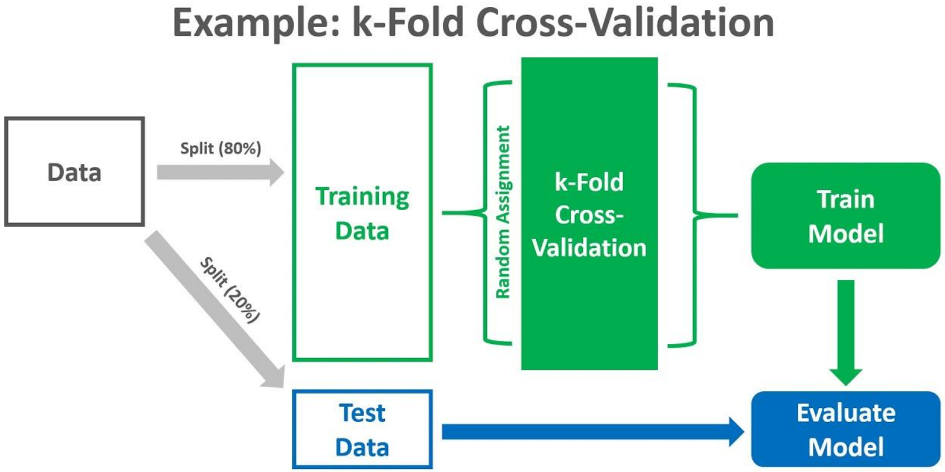
- Y, our target variable, is ‘Streams’ and X is all integer predictor variables
- Training set is 80% of of our original data
- Test set is 20% of our original data



Cross Validation

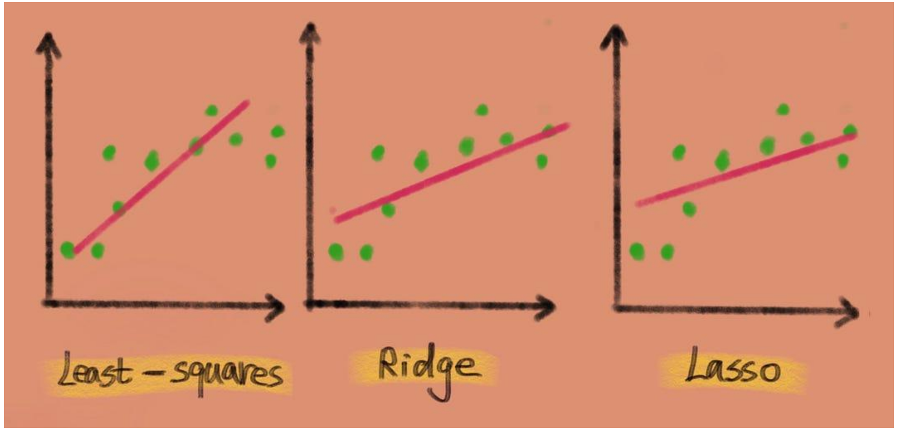
- Used K Fold function to define 10 cross validation folds

- Training data is divided into 10 folds. The model is trained and evaluated 10 times, using a different fold as the validation set each time



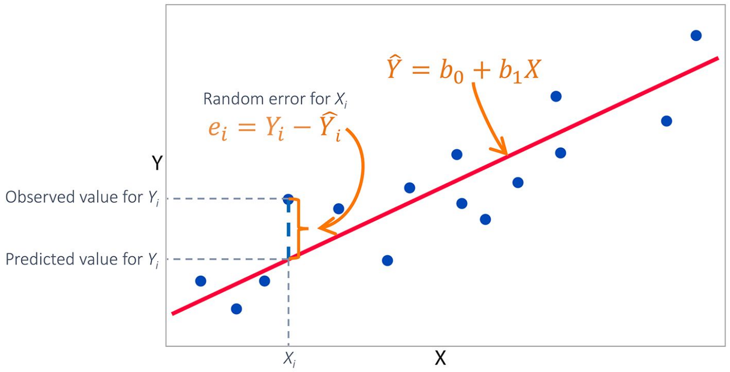
Benchmarks

To evaluate the performance of the PCR model, we run three baseline models (Standard Linear Regression, Lasso Regression, and Ridge Regression) and save the RMSE scores.



Linear Regression

- Use least-squares to fit a line to the data
- Sum up the squared residuals
- Find the rotation with the “least squares”



- **lin_reg = LinearRegression().fit(X_train_scaled, y_train)**

Lasso Regression

- Least squares + Lambda(|Slope|)
  - Least Squares: minimized sum of the squared residuals
  - Lambda is determined by cross validation

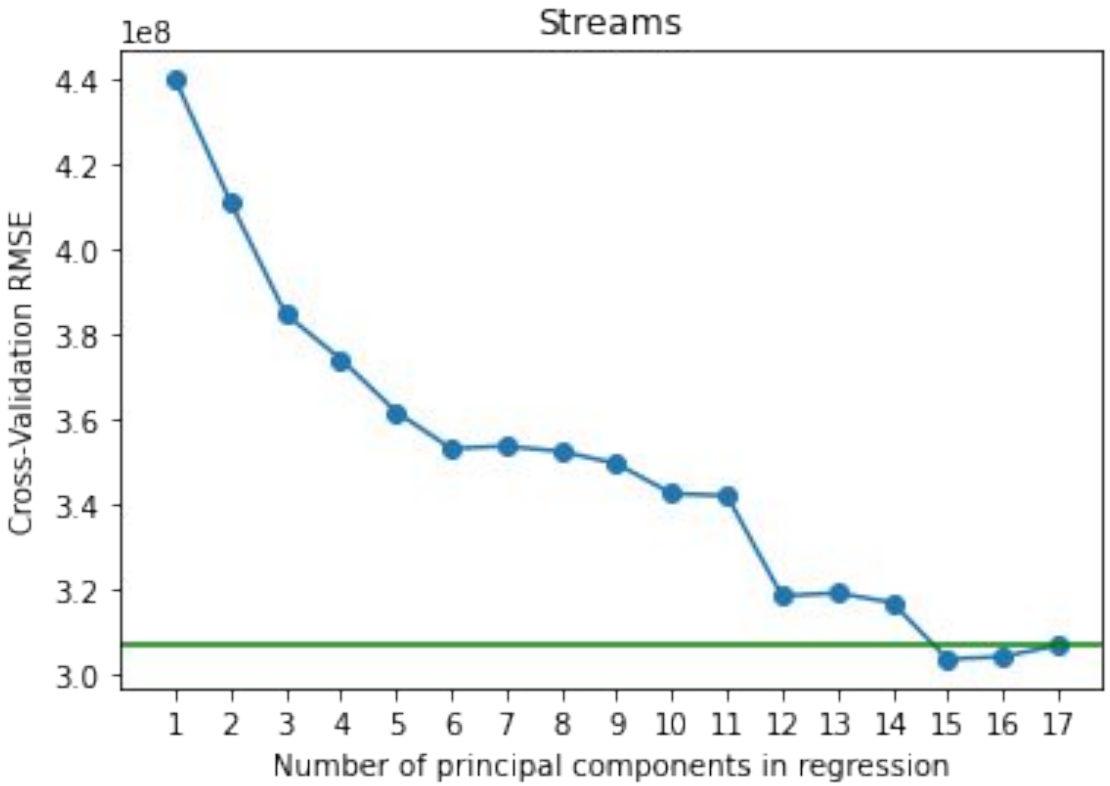
- **lasso_reg = LassoCV().fit(X_train_scaled, y_train)**

Ridge Regression

- Least squares + the “Ridge Regression Penalty”
  - Least Squares: minimized sum of the squared residuals
  - Ridge Regression Penalty: Lamba + slope^2

- **ridge_reg = RidgeCV().fit(X_train_scaled, y_train)**

RMSE vs Number of Principal Components



Training set performance of PCR improves (RMSE decreases) with more principal components. Lowest RMSE is with 15 principal components

RMSE Train Set

	RMSE (Train Set)
Linear Regression	3.066110e+08
Lasso Regression	3.063346e+08
Ridge Regression	3.061236e+08
PCR (15 components)	3.493991e+08

RMSE Test Set

	RMSE (Test Set)
Linear Regression	2.927741e+08
Lasso Regression	2.905677e+08
Ridge Regression	2.910445e+08
PCR (15 components)	3.142983e+08

Why are the RMSEs so high?

- RMSE measures the average difference between values predicted by a model and the actual values
  - High RMSE means large distance between predicted and actual values
- Multicollinearity
  - It’s possible the spotify playlist and apple music playlist variables are correlated so the principal components may inherit these issues
- Streams data points are separated far from each other

Works Cited

Karunakaran, Dhanoop. “Principal Component Analysis(Pca).” *Medium*, Intro to Artificial Intelligence, 10 July 2023, medium.com/intro-to-artificial-intelligence/principal-component-analysis-pca-cd282196b7d5.

Desai, Utsav. “Mastering Dimensionality Reduction: Exploring PCA and SVD Methods.” *Medium*, Medium, 3 May 2023, utsavdesai26.medium.com/mastering-dimensionality-reduction-exploring-pca-and-svd-methods-f7d55c9ca3c9.

Leung, Kenneth. “Principal Component Regression - Clearly Explained and Implemented.” Medium, Towards Data Science, 14 Sept. 2022 towardsdatascience.com/principal-component-regression-clearly-explained-and-implemented-608471530a2f.

https://github.com/kennethleungty/Principal-Component-Regression/blob/main/notebooks/Principal_Component_Regression_Wine_Quality.ipynb

https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#ridge-and-lasso-regression