Research Experience for Undergraduates (REU) in Data-Enabled Industrial Mathematics

Comparison of Dimensionality Reduction Techniques for Prediction of Plutonium(IV) in Nitric Acid Concentrations

REU Director: Dr. Mihhail Berezovski - *Embry-Riddle Aeronautical University, Daytona Beach, FL*

Industrial Partner: Dr. Aaron Luttman - *Pacific Northwest National Laboratory, Richland WA*

Authors

Zachariah Kline
Emily Armstrong
Jensen Bridges
Zoe Friedman
Kian Greene
Jacob Antici

Affiliations

Wisconsin Lutheran College
Assumption University
Oklahoma State University
Illinois State University
Embry-Riddle Aeronautical University
Arkansas State University

Background

At Pacific Northwest National Laboratory (PNNL) a team of chemists studying the nuclear fuel cycle collected data using ultraviolet-visible spectroscopy, a non-intrusive, analytical technique that measures the absorbance of particular wavelengths in solutions containing various amounts of Plutonium(IV) in nitric acid. These measurements are stored in a 1100x1650 data matrix, Y_Pu4_UVvis, where there are 1100 measurements of Pu(IV) in nitric acid taken at 1650 different wavelengths.

Objective

With this data, we will perform dimensionality reduction techniques to be able to preserve trends in the original data as we represent it in a manner that we can visualize, namely principal component analysis (PCA) and non-negative matrix factorization (NMF). Based on trends we observed in the data, we formed prediction models such that given an unknown sample's absorption, we can predict the sample's concentration of Pu(IV) and nitric acid.

Data

Since Y_Pu4_UVvis has dimension 1100x1650, we have over a million absorption values to work with. The 1100 rows consist of 110 distinct Pu(IV) in nitric acid concentrations; the chemists reproduced each distinct sample 10 times to account for human error. The 1650 columns represent the wavelengths [ranging from ultraviolet-visible light waves [380-1080 nanometers] at which absorption of lightwaves was measured. Visualization was the first step in formulating our approach, **Figure 1** plots the absorption rates averaged each set of 10 replicates for all 1650 wavelengths. Each distinct replicate is a distinct colored curve on the graph.

Implementation

Both of the following methods will be used to reduce our data into smaller dimensions, so we can more easily visualize relationshtrends:

- Principal Component Analysis (PCA)
  - PCA reduces the data based on the directions that display the most variance; data points are projected onto the derived principal components in order to see their representation for a chosen amount of components or dimensions.
- Non-negative Matrix Factorization (NMF)
  - NMF, as the name suggests, uses two smaller matrix factors that approximate the original matrix. Since Y_Pu4_UVvis has dimension 1100 x 1650, our two smaller matrices, W and H, will have dimensions 1100 x r and r x 1650 respectively, where r is our chosen rank or dimension.

Related Literature

Lines, A. M., Adami, S. R., Sinkov, S. I., Lumetta, G. J., & Bryan, S. A. (2017). Multivariate analysis for quantification of plutonium(iv) in nitric acid based on absorption spectra. Analytical Chemistry, 89(17), 9354–9359. https://doi.org/10.1021/acs.analchem.7b02161

Results: Principal Component Analysis [PCA]

To begin PCA, we centered the data by taking the mean of Y_Pu4_UVvis and subtracting it from each index of the matrix. Next, the covariance matrix must be calculated to determine how variables are interrelated. The eigenvectors and eigenvalues are derived from the covariance matrix. The eigenvectors are ordered and point in the direction of greatest variance in the data; their corresponding eigenvalues are magnitudes of the vectors. Together, these corresponding values give us our ordered principal components (PC); from our 1650 possibilities, we must choose n principal components onto which we will project Y_Pu4_UVvis.

Using Python

In python, we imported PCA as a function from sklearn.decomposition. and used this to derive our principal components. In the direction of greatest variance of the data, PC1 accounts for 89.41% and likewise, PC2 shows 9.32% of the variance. It is important to note that the first two components are orthogonal. In Figure 2 the 2 dimensional representation of where n=2, our data is projected onto axes PC1 and PC2

Beer Lambert Law

By examining Figure 2, we see a direct correlation between both concentrations Pu(IV) and nitric acid and absorbance rates. This is expected by the statement of Beer-Lambert Law: a direct relationship exists between absorbance and concentration; the path length through the sample and the concentration of the Pu(IV) in nitric acid are proportional to ultraviolet-visible light absorbance.

This direct proportionality inspired our method of forming a prediction model- linear regression. Before we can discuss how a linear regression model performs, we must explain how we trained our model.

Train-Test Split

We split the data into different uses for creating our predictive model with a portion to train our model and the rest to be used as data to test its performance. By separating test data from training, we ensure that the model has not seen the absorbance data to which is it trying to assign to a particular Pu(IV) in nitric acid concentration.

Random Sampling

The two main methods for sampling we tried were stratified and random sampling. Stratified sampling aims to draw conclusions based on disjoint subgroups within the entire dataset; each 10 replicates are considered a strata of which we have 110. Using stratified sampling and taking an 80% train and 20% test split is not ideal because 8 of every 10 replicates are used to train a model. If the test data consists of 2 replicates of which the model has already seen 8, these "predictions" do not display true accuracy of the model. For this reason,, we use random sampling.

Our Prediction Model

Taking into consideration how our models performs based on varying number of principal components, differing amounts of train-test data and type of sampling helped us decide on our model. We will review the main measures of accuracy for choosing how to train our model: mean squared error (MSE) [**Figure 4**] and R^2score [**Figure 3**]. Minimizing MSE improves accuracy, and an R^2 close to 1 indicates success of our predictions. 27 principal components allow our model to predict both concentrations with $R^2 = .99$

Results: Non-negative Matrix Factorization [NMF]

NMF is a machine learning algorithm that separates a large m x n matrix into two smaller, non-negative matrices W and H. In order to approximate Y_Pu4_UVvis, W must have dimension 1100 x r and H must have dimension r x 1650. The rank, r, determines what dimension we are reducing our data into. Similar to PCA, choosing rank will affect the MSE and R measuring our models prediction. For comparability, we will continue to use the some of the same values from PCA such as train-test split, random sampling, and linear regression.

Our Prediction Model

Using linear regression on our W matrix, we created a model; Figure 5 shows how MSE for Pu(IV) and nitric acid respond to different ranks [Figure 5]. Again, Pu(IV) is more easily predicted in lower dimensions, but due to nitric acid's relatively high error, we will use a rank that will suit both Pu(IV) and nitric acid.

Conclusion:

Both PCA and NMF were successful in reducing the dimension of our dataset in order to better understand trends within the data. PCA was the process that we began our research on; it is popularly used in spectroscopy, but it does yield negative numbers in its calculations; this is problematic since this chemistry problem deals with concentrations and absorbances which will never be negative. For this motivation, we expanded our research to non-negative matrix factorization. Both PCA and NMF performed well in predicting concentrations of an unknown sample based on its ultraviolet-visible absorbance, however, NMF has the advantage of only using non-negative values. Overall, both methods were able to form an accurate prediction model on dimensionality reduced data.

Support for the program has been provided by the National Science Foundation (NSF) through REU Award Number DMS - 2050754.

Graphs for PCA

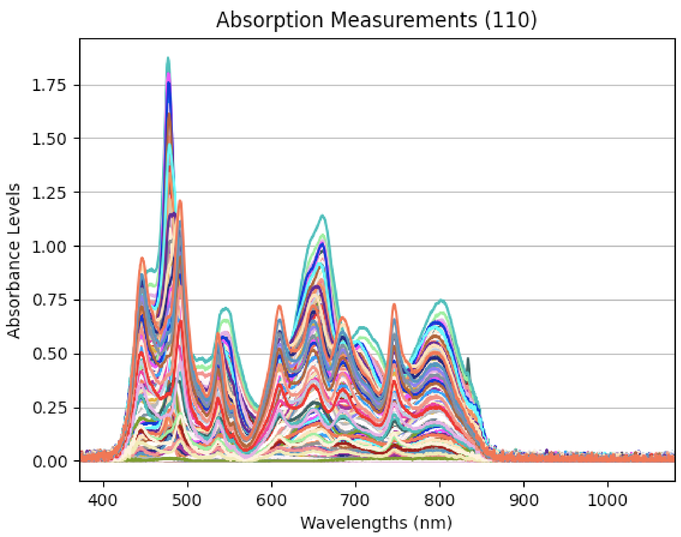

Figure 1- *Each curve represents a solution containing a distinct amount of Pu(IV) in nitric acid. Each of the 110 different colors represent the average absorbance of a set of replicates*

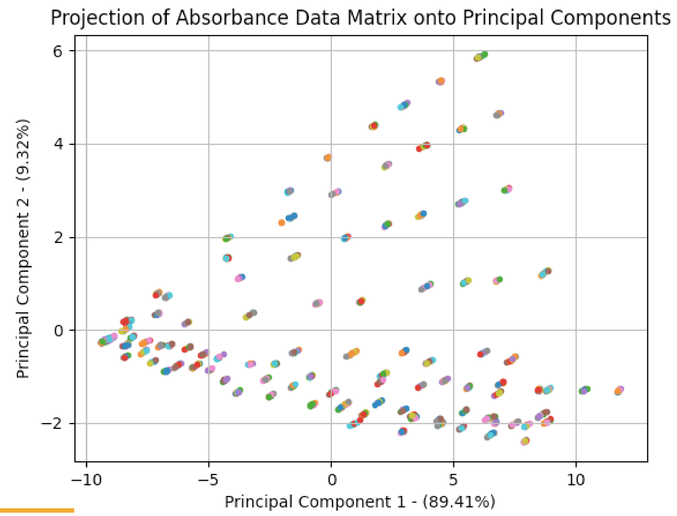

Figure 2- *The absorbance data is transformed into 2 dimensions by projection onto the first two principal components. The data organizes itself along 11 distinct lines of positive slope. Each of these 11 lines represent a distinct nitric acid concentrations; as you move left to right within each line, each pf the 10 clusters in each line represent increasing Pu(IV) concentrations.*

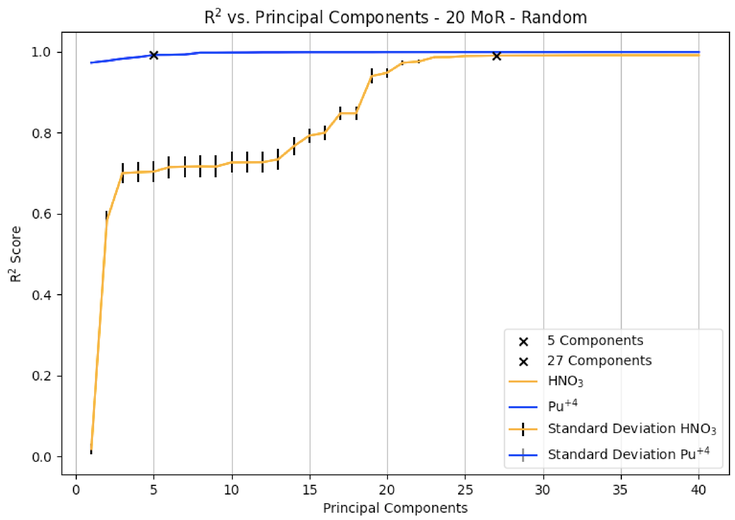

Figure 3- *R gives us a percentage of the number of predicted points that fall onto our line of best fit and how it changes with the number of principal components, n. As we can see, Pu(IV) is more easily predicted accurately with less components, but we will focus on the minimum number of components needed to predict both with R = .99; this accuracy requires 27 components.*

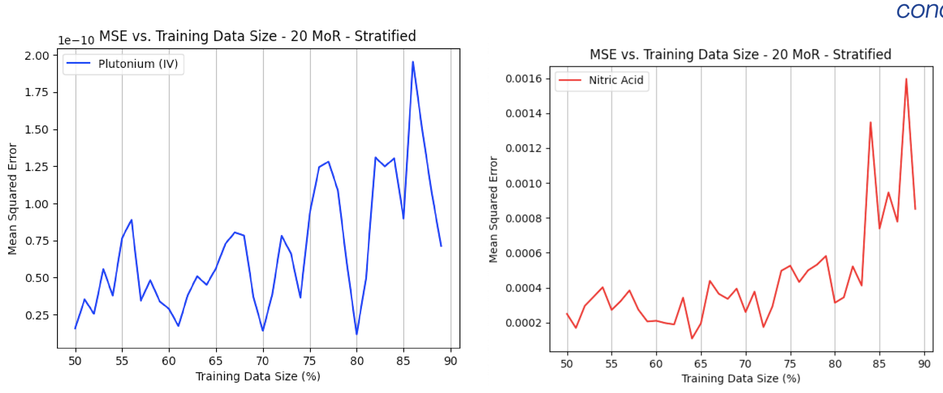

Figure 4- *The MSE of our prediction model's ability to predict Pu(IV) [on the left] and nitric acid [on the right] are plotted against different training data sizes. We hope to minimize MSE by choosing a sufficient percentage of training data . From these visuals, we decided to use 80% of randomly sampled data to train our model.*

Graphs for NMF

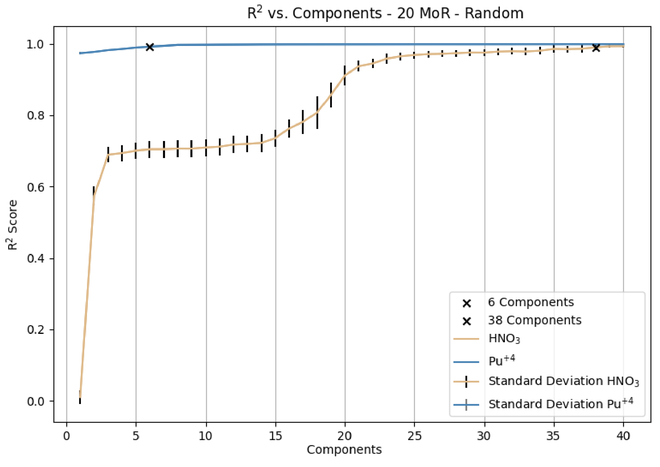

Figure 5- *This graph shows both Pu(IV) and nitric acid's accuracy of predictiions as rank varies. As our rank goes above 30, the R is very close to 1. With our goal being able to predict both concentrations with R. =.99, we must used 38 as our rank.*

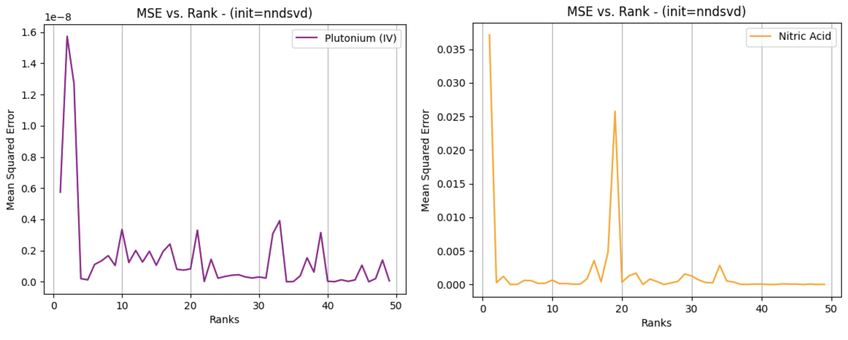

Figure 6- *The graphs above measure how the MSE of the linear regression prediction model changes with respect to rank. As we look to minimize MSE, a rank greater than 20 will keep our MSE for Pu(IV) predictions quite low, but nitric acid is not as easy to predict. Based on the two curves, we will use 37 as our rank.*