Sanjeev Chauhan:

**Questions** (answer all 3)

1. Given the bias-variance tradeoff, how do you evaluate the role using a PCA has in selecting features for your regression model? To be more precise, imagine you have 10 independent variables and a two dimensional latent space captures 70% of the total variance. Will using these latent variables improve your regression model’s ability to generalize out-of-sample? Or are there possible downsides? Which variables would you include in the PCA and which would you include separately in the regression? Finally, if you are running a purely predictive model, why or why not would you use feature selection of this kind?

**Given the bias-variance tradeoff, how do you evaluate the role using a PCA has in selecting features for your regression model?:**

So first what is the bias- variance tradeoff:

Random error is made up from bias and irreducible error. Bias can be theoretically reduced so the random error is just the irreducible error. In other words *Error = bias + variance*. Holding the error fixed yeads to an inverse relationship between bias and variance. High variance could lead to overfitting the dataset and poor results on new data. A model with high variance won’t work well at representing a new dataset as it is likely overfit to the training data. High bias could show consistent results between datasets but underfits the data. A model with High bias would work consistently between datasets but would not produce a result accurate to the data it receives. A naïve model has high bias.

We want a model that has a balance between bias and variance, It should be able to work on other datasets and still be responsive to difference in the datasets.

PCA modifies the variables used in your dataset to help lower variance and combat overfitting. PCA reduces how many dimensions your model has and less dimensions make a model harder to overfit and easier to understand by humans. PCA also helps reduce the amount of colinear variables. We want our IVs to be independent from each other as we are trying to find the effects of each variable onto our y. If we determine a bunch of different variables are related to each other (colinear). Then we can all those IVs and combine them into a new PCA variable. This reduces the dimensions of our dataset from how many IVs we put into our PCA into how many PCA variables we end up using.

Diagram

Description automatically generated

Here we can see the use of PCA to reduce 3 dimensions into 2. A linear regression can easily be done in a 2d space but is much harder to do in a 3d space.

I see PCA as a very useful way to prevent overfitting and reduce variance. It helps to reduce dimensions and combine colinear variables.

**Will using these latent variables improve your regression model’s ability to generalize out-of-sample?:**

Yes, the main problem with a model not being able to generalize out-of-sample is overfitting. Overfitting is from high variance in the model. Overfitting happens when a model pays attention to small changes too much. An overfit model would work really well for the dataset it is trained on but not work well on any other dataset. A way to counteract overfitting is to use latent variables.

Latent variables remove some small part of variance when IVs are combined to form a latent variable. This combined with lower dimensions when using latent variables (as each latent variable contains multiple IVs) makes it harder for the model to overfit.

Using a two dimensional latent space that captures 70% of the variance instead of ten IVs would still allow the model to account for a good amount of variance in the dataset with using way less dimensions. Using less dimension in the dataset would allow it to generalize out of sample better.

**Or are there possible downsides?:**

There are downsides. If latent variables are chosen poorly, then important variables could be removed.

In our case, maybe a few of the 10 IVs correlate strongly to our y and the others aren’t useful. By putting all the variables into our Latent dimensions we are hurting our model by including non-useful variables. It would be better to drop Ivs not being used by our model first before making a latent space.

**Which variables would you include in the PCA and which would you include separately in the regression?:**

I would group variables that show correlation into a PCA and use a latent variable instead to lower dimensionality. I would first make a scatter plot of my data to spot correlation between different variables. I would include variables that either don’t show any correlation to other variables in the data or variables that have a strong correlation with Y separately in the regression.

if you are running a purely predictive model, why or why not would you use feature selection of this kind?

I would not. The ability to predict Y matters the most and not what goes into the model. I would just use a decision tree like Xgboost / random forest instead. Doing PCA is helpful when your trying to tell a story between the IVs and prediction. But If you don’t need to do that then running a decision tree is much easier.

1. Imagine you are trying to run a campaign for a presidential candidate. In the US, these campaigns compete in each state and it is winner-take-all for the candidate that wins each state. If you wanted to build a predictive model of how each candidate would do and identify factors that would help you advise your candidate, what unit of analysis would you focus on (i.e., what does a row in your dataset look like)? What challenges to inference exist, especially with respect to strategic behavior? What IVs would you collect?

If you were trying to build a predictive model of how each candidate would do in a presidential campaign, the unit of analysis would likely be the state. Each row in the dataset would represent a different state, and the variables in the dataset would be the independent variables (IVs) that might influence the outcome of the election in that state.

There are several challenges to inference that would need to be considered when building this type of model. One challenge is strategic behavior, which refers to the fact that candidates and their campaigns may alter their behavior in response to the actions of their opponents. For example, a candidate may choose to allocate more resources (e.g., money, staff, advertisements) to a particular state if they believe they are in danger of losing there. Also a candidate would pay attention to “Swing States” to see if they have an opportunity to take over a state that normally swings the other day. This can make it difficult to disentangle the causal effects of the independent variables on the outcome, as the outcome may be influenced not only by the IVs, but also by the strategic behavior of the candidates.

In terms of IVs, there are many factors that might influence the outcome of a presidential election in a particular state. Some potential IVs to consider might include:

* Demographic characteristics of the state, such as the age, race, and education level of the population
* Economic indicators, such as the unemployment rate, median income, and GDP growth rate
* Political factors, such as the party affiliation of the state's elected officials and the state's history of voting for particular parties
* Social and cultural factors, such as the state's attitudes towards particular issues (e.g., gun control, abortion)
* Campaign activities, such as the amount of money spent on advertising and the number of campaign events held in the state

In a dataset each row would be a state with the Y variable being which way the state swings and the IV’s being the ones listed above.

3. For a given sample, you start with a purely linear regression model and then try a polynomial regression of order 3. In both cases, model fit is similar using MSE and both models show heteroskedasticity. Which models would you prefer, all else equal? Given the presence of heteroskedastic errors, what do you assume is wrong with your approach? How can you correct this problem? Finally, let’s say you try a decision tree approach and MSE improves dramatically. What would you infer about the relationship of y ~ *f*(**x**)?

It is generally preferred to use the model with the lower MSE, but in this case the MSE are similar. First off, I would choose the simpler model. (linear regression). The simpler model would be easier to interpret and overall, less prone to overfitting. However, it is important to consider other factors as well, such as the interpretability and complexity of the model, and the sample size and number of predictor variables.

Linear regression is a simple and widely used approach that assumes a linear relationship between the predictor variables and the response variable. It is easy to interpret and implement, and works well with large sample sizes. However, it may not be able to capture more complex relationships between the predictor(s) and response.

Polynomial regression is a more flexible approach that allows for more complex relationships between the predictor(s) and response by adding additional powers of the predictor variables as additional features. It can be more powerful than linear regression in some cases, but it can also be more prone to overfitting, particularly with small sample sizes or when the degree of the polynomial is too high. It is also generally more difficult to interpret than linear regression.

Ultimately, the choice between linear regression and polynomial regression will depend on the specific characteristics of your data and the goals of your analysis.

The presence of heteroskedastic errors in the linear and polynomial regression models suggests that the variance of the errors is not constant across the range of the predictor variable(s). This can be caused by several factors, such as nonlinear relationships between the predictor(s) and the response, or the presence of outliers in the data.

One way to correct for heteroskedasticity is to transform the response variable and/or the predictor variables in a way that stabilizes the variance of the errors. For example, taking the square root or logarithm of the response variable can often help to stabilize the variance. Another approach is to use weighted least squares regression, in which the errors are weighted differently depending on their size.

If the MSE improves dramatically when using a decision tree approach, it could indicate that the relationship between the response variable and predictor variables is non-linear. Decision trees are well-suited for modeling non-linear relationships, so this improvement in MSE could be an indication that the relationship between y and f(x) is better represented by a non-linear model. However, it is important to carefully evaluate the decision tree model to ensure that it is not overfitting the data.

**Data problem** (mandatory)

With the attached data, your DV is modern day inequality (measured by gini\_disp; see [https://en.wikipedia.org/wiki/Gini\_coefficient)](https://en.wikipedia.org/wiki/Gini_coefficient) and your IV’s are various measures of countries at different points in time. Your sample is small b/c there are only so many countries in the world. Turn in your “best” model and a brief explanation of why you did what you did. Variables are as follows:

|  |  |
| --- | --- |
| Ygini\_disp | DV on inequality |
| country | Country name |
| federalism\_GT | federalism variable |
| id | Country ID |
| region\_wb | regional dummy |
| gdp | gdp |
| statehiste1500\_02n | state agricultural history at 1500 AD |
| origtime2 | origin time of state |
| eleva | elevation |
| avg\_temp | average temp |
| Maddison\_gdppc\_1990\_estimate\_ln | gdp / capita in 1990 |
| lp\_lat\_abst\_fill | latitude |
| mountains |  |
| log\_ocdistance\_new | distance from center of country to  ocean |
| rugged |  |
| tropical |  |

|  |  |
| --- | --- |
| pmean | preciptiation mean |
| irri\_impact5 | impact of irrigation |
| frstdays | frost days |
| sd\_emeanclip | variance in elevation |
| Urbanpopulationoftotalpop |  |
| dist2suitable\_km\_new | distance from center to port |
| Fixedtelephonesubscriptionsp |  |
| Employmentinagricultureof |  |
| Accesstoelectricityofpopu |  |
| pln\_sxHr\_mean | plantation crop suitability |
| agyears\_ext | length of time using advanced  agriculture |
| popd\_1500AD | population at 1500 AD |