Update: Total PCA v. OLS

Art Tay

Table of Contents

Theory

Background

OLS Refresher

Let X be a $n \times p$ centered design matrix, and Y be the $n \times 1$ response vector. Then the projection matrix H that projects Y to $\hat{Y}$, the closest value to Y in the $col(X)$ is given by

$$\hat{Y} = HY = X(X^T X)^{-1} X^T Y$$

PCA Refresher

Since $X^T X$ is symmetric, it can be decomposed into PDP^T, where P is the $p \times p$ orthonormal matrix of the eigenvectors of $X^T X$.

- This means that P here is equivalent to the "PCA rotation matrix".

Intuition

- We will define a "Total PCA Transformation" to be $Z = XP$.
- Z is a $n \times p$ matrix that is a linear transformation of X. Therefore $col(Z) = col(X)$.
- Since the predictor space has just been rotated, the span hasn't changed and $\hat{Y}$ should still be the closed point to Y in the new predictor space Z.

Proof

- Let H' be the projection matrix that projects Y onto Z.

$$\begin{aligned} H' &= Z(Z^T Z)^{-1} Z^T && (1)\\ &= XP((XP)^T XP)^{-1}(XP)^T && (2)\\ &= XP(P^T X^T XP)^{-1} P^T X^T && (3)\\ &= XPP^T (X^T X)^{-1} PP^T X^T && (4)\\ &= X(X^T X)^{-1} X^T && (5) \end{aligned}$$

- Thus $\hat{Y}_{PCA} = \hat{Y}_{OLS} \Rightarrow$ the Training and Testing errors must be identical between the two methods.

Code

Problem

- Old Code:

```
pca_recipe <- recipe(Y ~ .,
    data = dummy_dataset) %>%
    step_pca(all_numeric_predictors(),
        threshold = 1)
```

- New Code:

```
pca_recipe <- recipe(Y ~ .,
    data = dummy_dataset) %>%
    step_pca(all_numeric_predictors(),
        num_comp = 30)
```

Potential Problem

What could this mean?

- There are some samples where certain principal components explain 0 additional variance.
- Those principal components have a corresponding eigenvalue of 0.
- $det(X^T X) = 0 \iff (X^T X)^{-1}$ does not exist $\iff$ Perfect Collinearity.
- The model will still "fit", but it will be uninterpretable.

Toy Example

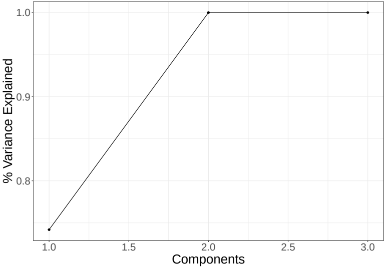
- Suppose a true data generating model:

$$Y = X + Z + \epsilon$$

where X, Z, ϵ are all $\sim N(0,1)$.

- But suppose a third predictor, $W = X + \frac{1}{2}Z$ is included.

Toy Example



models	MSE	RMSE
OLS	0.7579	0.8706
PC1-2	0.7579	0.8706
PC1-3	0.7165	0.8465

- Maintaining all components under perfect correlation can “trick” the algorithm into using useless information, because it will think its an uncorrelated predictor.

Actual Problem

Unexpected Behavior

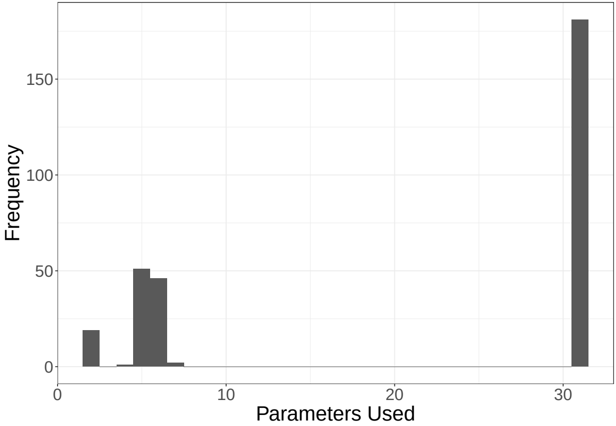


Figure 1: Number of Components Retained by Total PCA Models Under the Old Code

Appendix

Low Correlation Case

Table 2: Modeling results under low correlation between parameters

Model	Training RMSE	Test RMSE	Parameters
OLS	(0.81, 0.85)	(1.19, 1.2)	(19.49, 21.57)
LASSO	(0.82, 0.86)	(1.17, 1.18)	(15.13, 15.81)
PCA	(0.81, 0.85)	(1.19, 1.2)	(25.91, 26.93)
PCA + Cutoff	(1.88, 1.99)	(2.16, 2.19)	(9.55, 10.07)
PCA + LASSO	(0.84, 0.89)	(1.19, 1.2)	(13.97, 15.55)
PLS	(0.81, 0.85)	(1.19, 1.2)	(13.15, 14.43)
PLS + LASSO	(0.85, 0.89)	(1.17, 1.18)	(14, 17.74)

[1] 99% mean t confidence intervals.

[2] Parameters means non-zero for LASSO type models and significant at $\alpha = 0.05$ otherwise.

Moderate Correlation Case

Table 3: Modeling results under moderate correlation between parameters

Model	Training RMSE	Test RMSE	Parameters
OLS	(0.8, 0.83)	(1.18, 1.19)	(14.99, 19.69)
LASSO	(0.83, 0.88)	(1.13, 1.14)	(12.86, 14.08)
PCA	(0.8, 0.83)	(1.18, 1.19)	(24.51, 25.53)
PCA + Cutoff	(1.77, 1.86)	(1.91, 1.94)	(4.9, 5.32)
PCA + LASSO	(0.85, 0.9)	(1.15, 1.16)	(12.83, 14.37)
PLS	(0.8, 0.83)	(1.18, 1.19)	(15.79, 17.13)
PLS + LASSO	(0.85, 0.89)	(1.12, 1.13)	(15.16, 17.66)

[1] 99% mean t confidence intervals.

[2] Parameters means non-zero for LASSO type models and significant at $\alpha = 0.05$ otherwise.

High Correlation Case

Table 4: Modeling results under high correlation between parameters

Model	Training RMSE	Test RMSE	Parameters
OLS	(0.8, 0.84)	(1.23, 1.25)	(16.35, 19.11)
LASSO	(0.86, 0.91)	(1.15, 1.17)	(9.27, 10.59)
PCA	(0.8, 0.84)	(1.23, 1.25)	(20.8, 22.1)
PCA + Cutoff	(1.7, 1.77)	(1.76, 1.78)	2
PCA + LASSO	(0.88, 0.91)	(1.16, 1.18)	(10.04, 11.44)
PLS	(0.8, 0.84)	(1.23, 1.25)	(15.26, 17.14)
PLS + LASSO	(0.87, 0.9)	(1.14, 1.15)	(14.16, 15.76)

[1] 99% mean t confidence intervals.

[2] Parameters means non-zero for LASSO type models and significant at $\alpha = 0.05$ otherwise.