Thinking about Principal Component Analysis and Lasso Regression

Sarah Ball

Principal Component Analysis on the MNIST dataset

The basic idea of Principal Component Analysis (PCA) is to reduce the dimensionality of the data. This is achieved by creating principal components (PC), whereby the first PC accounts for the largest possible variance of the data. The PCs are basically the directions of the axes where the most variance is captured. The vectors "consist" of eigenvalues, which in turn give an indication about the variance explained in each PC. We are now going to illustrate how PCA works visually on the MNIST dataset (LeCun et al., 2010). The MNIST dataset is a publicly available database of handwritten digits. The training set contains 60,000 examples, while the test set contains 10,000 examples.

If we apply PCA to the MNIST dataset, we see that almost the entire variance is explained by the first 100 principle components (Figure 1). Hence, one could think about reducing the dimensionality of the data by just using the first 100 principle components (PC) for further analysis.

Figure 1: Percentage of variance explained by each principle component

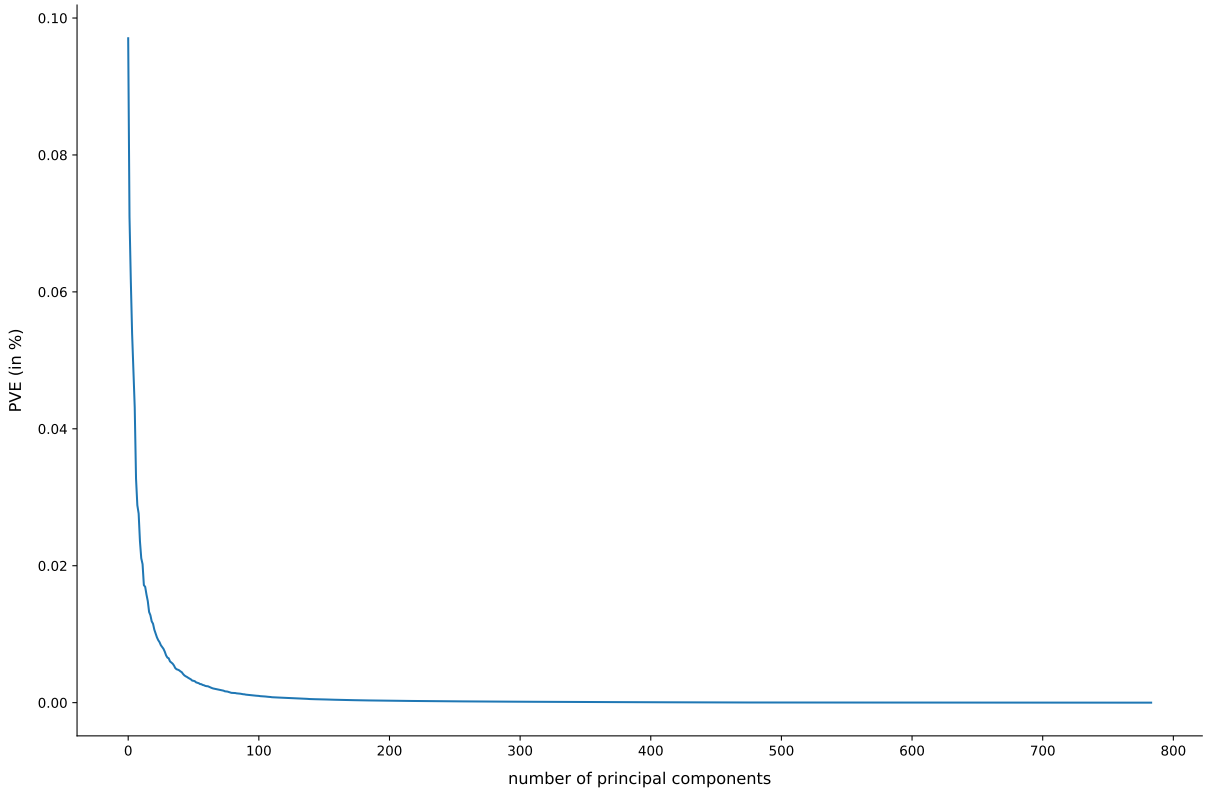


Figure 2 as well as Figure 3 illustrate how the PCA can aid at detecting different clusters in the dimensionality reduced plots (before we had 784 dimensions, which cannot be plotted). While the separation is not perfect, some structures emerge. When looking at the plot of the third and second principle component (Figure 4), the separation between the digit classes becomes less clear.

Figure 2: Number digits in dimensionality reduced space, PC1 and PC2

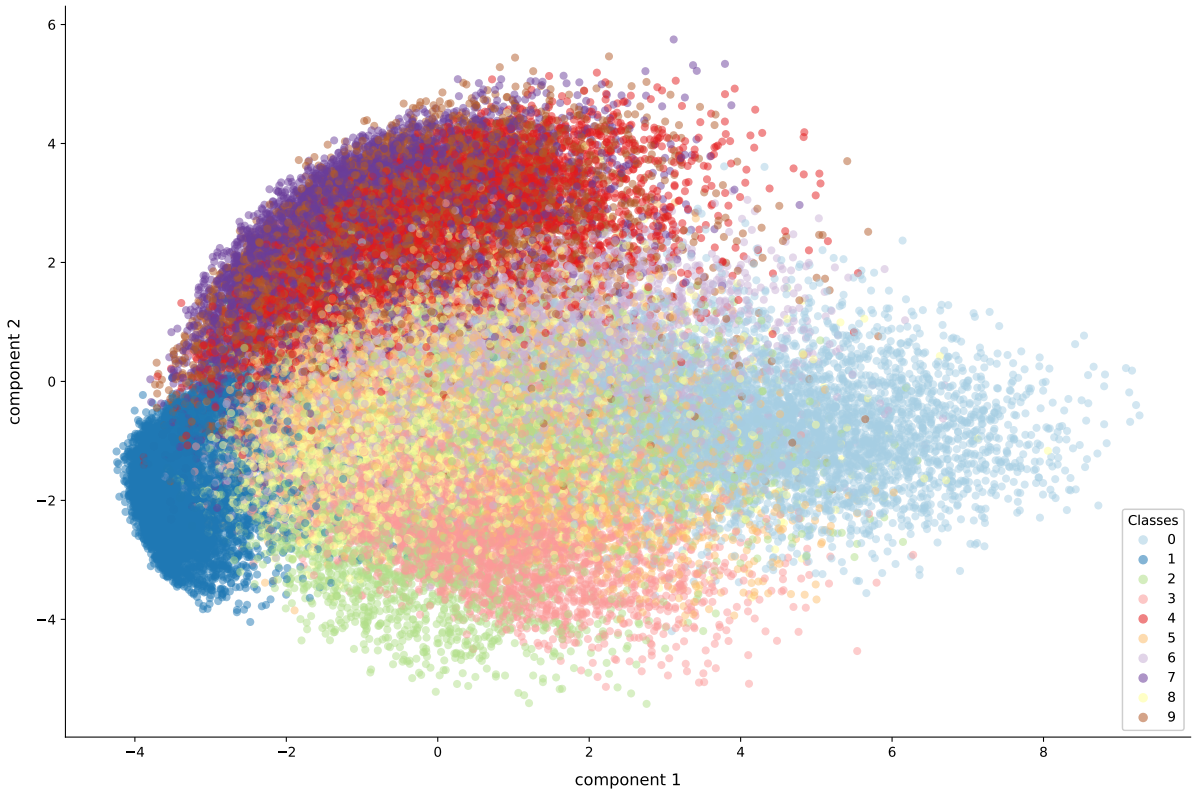


Figure 3: Number digits in dimensionality reduced space, PC1, PC2, and PC3

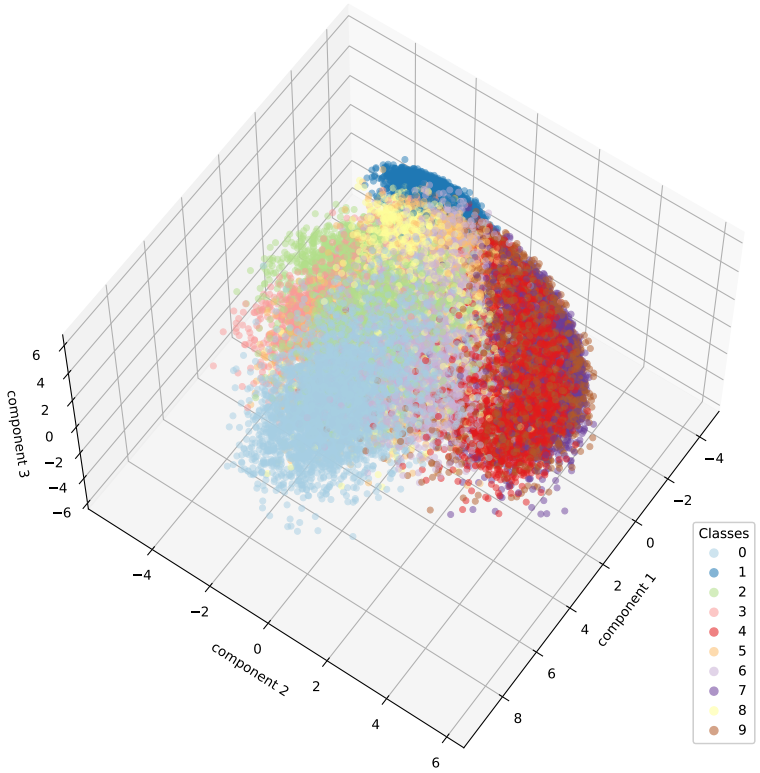
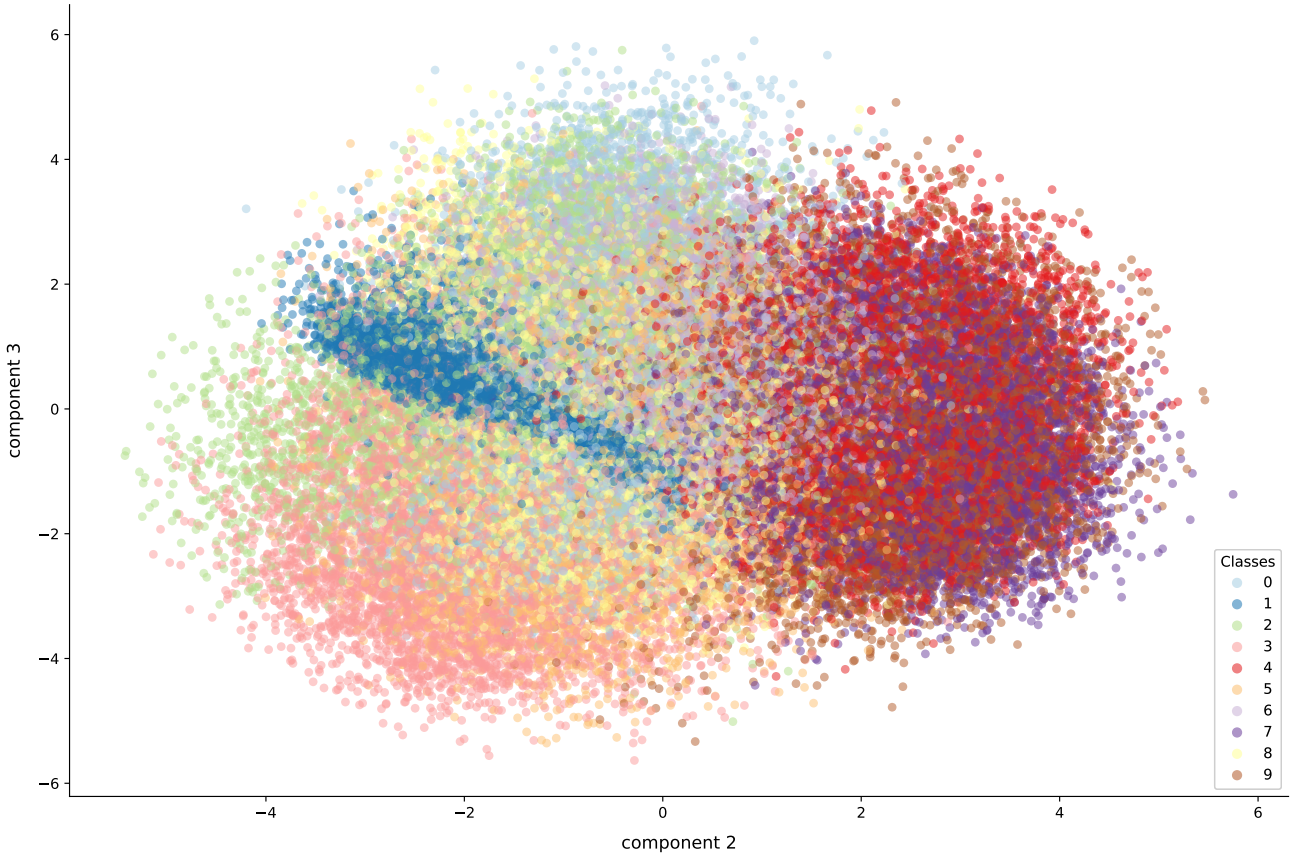


Figure 4: Number digits in dimensionality reduced space, PC2 and PC3



PCA and Lasso Regression

Next, we want to compare Lasso regression and regressions containing principal components. Why are we comparing those two? Both have the aim to reduce the dimensionality when conducting analyses.

Imagine the following scenario: You want to analyse the influence of 20 predictors on an outcome variable. Assume that these variables are correlated, and multivariate normally distributed with variance one.
Now, you add two more variables to your analysis, these two variables are standard normal and not correlated with any other predictor. A response variable Y was simulated in the following way:

$$Y = 0.5 + 0.1 * X_9 + 0.05 * X_{13} + 1.7 * X_{21} + 2.2 * X_{22} + 0.1 * \epsilon \quad (1)$$

Note that ϵ is standard normally distributed. You now want to conduct a regression analysis of Y on the 22 variables using either a PCA regression on the top two PCs or a Lasso regression using only two coefficients that are non-zero. You may now wonder which approach has the lower mean squared error (MSE). How can we answer this?

Let's begin with the PCA regression. For this approach you would first do a PCA including X_1 to X_{22}. Since the variables X_1 to X_{20} are correlated, the first PC will have eigenvalues which are higher for those variables than the eigenvalues for the uncorrelated X_{21} and X_{22}. This is because a lot of information about the entire data can already be explained by just capturing the variation of the correlated variables. Figure 5 illustrates this: X_{21} and X_{22} barely load on PC1. While their loadings are somewhat higher on PC2 (in absolut terms), other variables still contribute significantly more to PC2. For the regression, the first two PCs are used to transform the feature matrix, which basically results in a "downweighing" of X_{21} and X_{22}. Therefore, when performing a regression on the transformed feature values, the influence of X_{21} and X_{22} on Y is estimated too low in comparison to a Lasso regression without a previous PCA.
In the Lasso regression, only a *subset of features* is selected with the help of shrinking parameters with an $L1$ norm. In the scenario of the exercise, λ – the penalty parameter – is increased until only the coefficients of the two most influential variables on Y are non-zero. These are X_{21} and X_{22} (see Figure 6), as per the definition of Y those variables have the highest coefficients in the equation.

Figure 5: Loadings on first two principle components

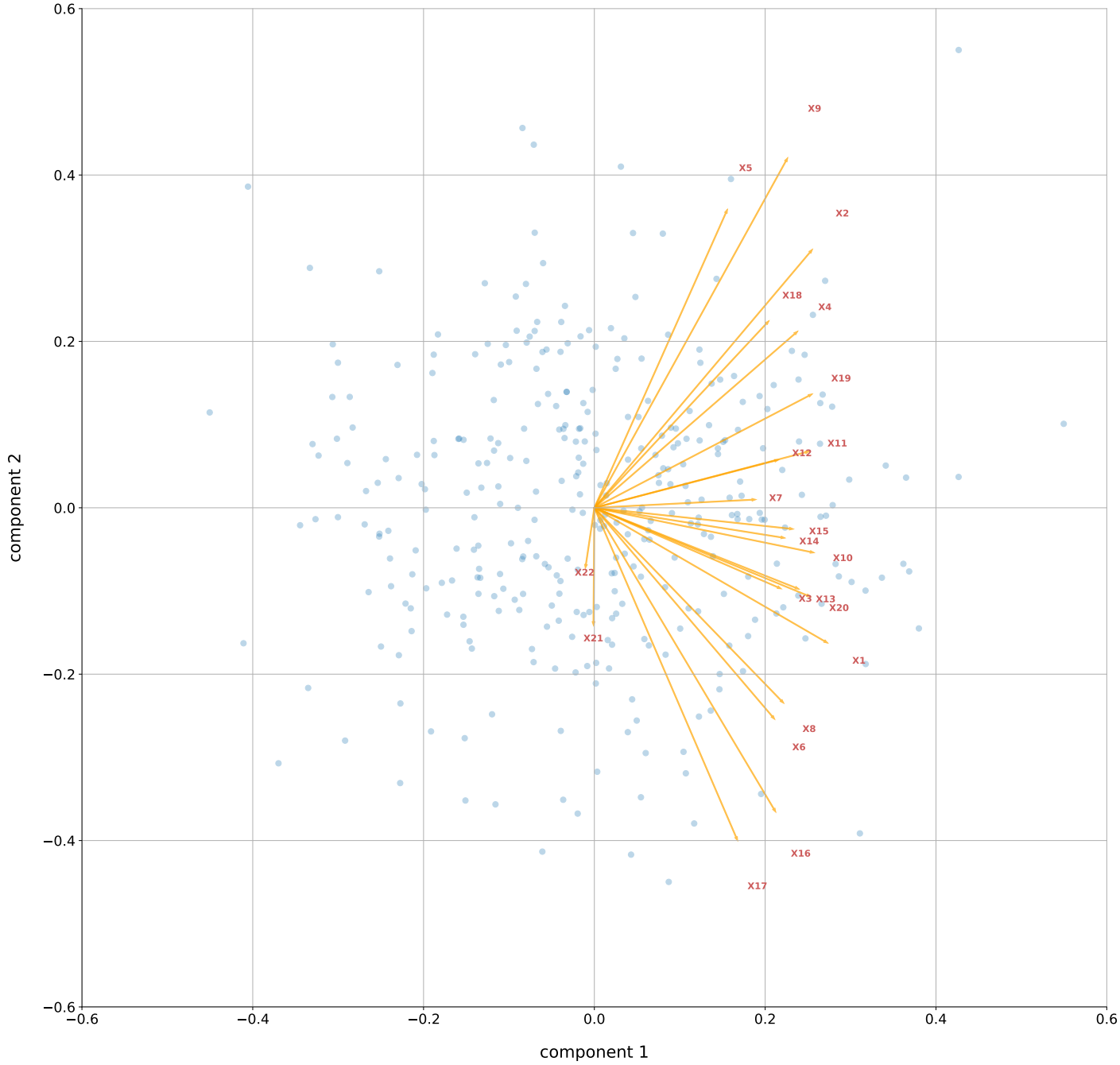
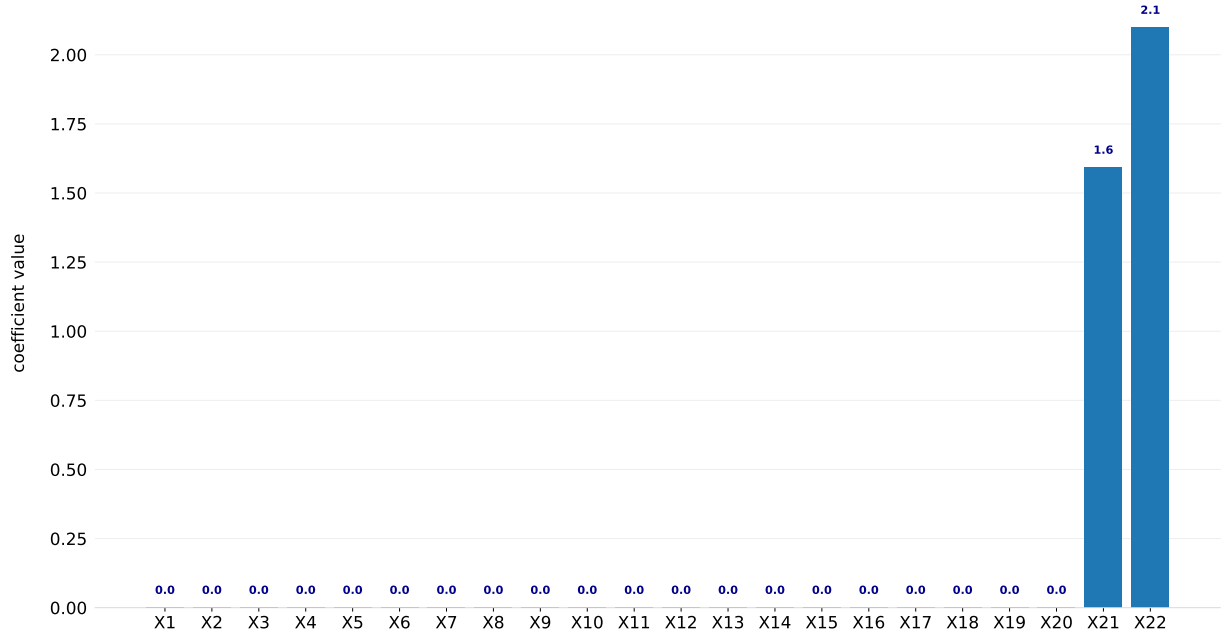


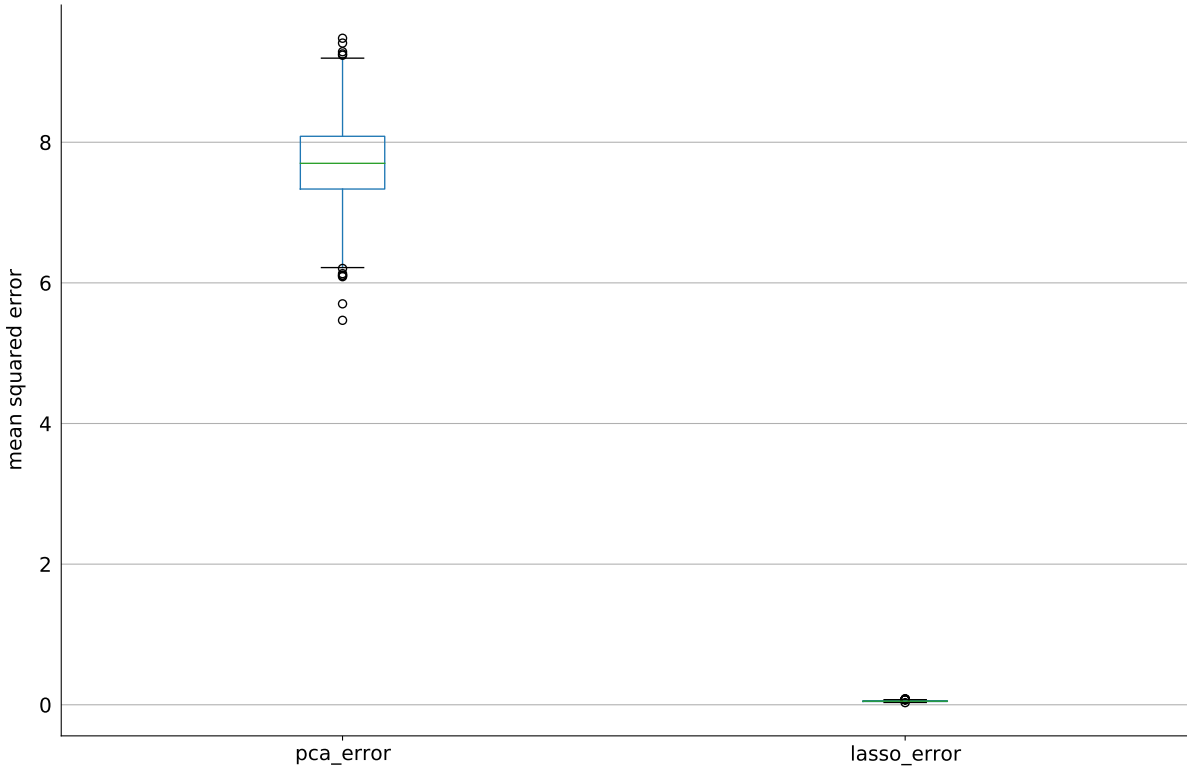
Figure 6: Coefficient values after lasso regression



Overall, the MSE for Lasso is lower than for the PCA regression. This is because while Lasso selects the two variables X_{21} and X_{22} as the most important features explaining the variance in Y, the influence of X_{21} and X_{22} is reduced by the PCA. Therefore, PCA regression performs worse.

To illustrate this further, we can simulate the two regressions with 1000 different random draws of 22 variables following the characteristics given in the description. The boxplots of the regressions' MSE in Figure 7 confirm our theoretical statement.
If the assumption of the exercise was that the PCA regression was only performed on the first 20 correlated variables without including X_{21} and X_{22}, the PCA regression would still have a higher MSE, because the influence of the most influential variables X_{21} and X_{22} is not captured in contrast to Lasso.
If the assumption of the exercise was that the PCA regression was only performed on the first 20 correlated variables and then the variables X_{21} and X_{22} were included for the regression, the PCA regression would perform better than Lasso. This is because in addition to the estimated influence of X_{21} and X_{22} we would have information on X9 and X13 via the first two PCs.

Figure 7: Boxplots mean squared error PCA and Lasso regression



Bibliography

LeCun, Y., Cortes, C., & Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, *2*.