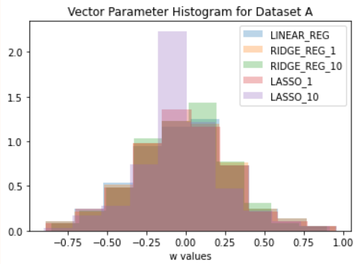
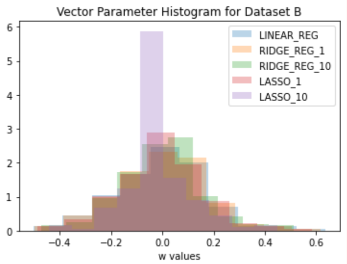
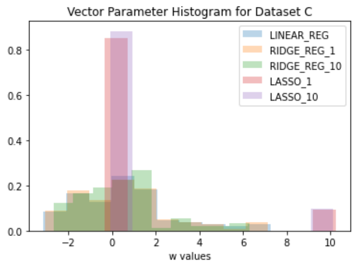
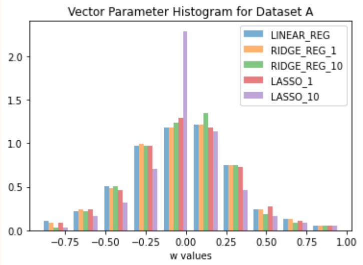
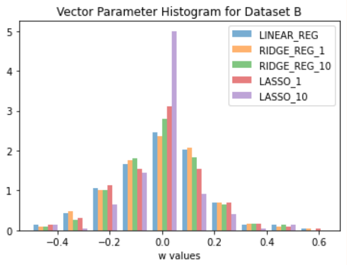
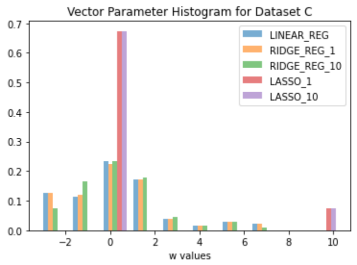
Answer 3.3

To throughly compare the lasso and ridge regression methods, we will have to divide these alpha parameters (1 and 10) by n for lasso in scikit-learn. This is mainly caused by the difference in objective function in scikit-learn. Specifically, the objective for the lasso incorporates the sample size n as the denominator of the mean squared error [1], whereas, the objective for the ridge regression does not take into account this term [2]. Hence, the regulation term α shall also divide by the sample size n in the lasso to match up with the ridge regression, so that both algorithm has similar regulation effects.

Test results are tabulated here:

Methods	Dataset A	Dataset B	Dataset C
Linear	MSE = 3.24740	MSE = 2.74268	MSE = 506.37271
Ridge 1	MSE = 3.13939	MSE = 2.61621	MSE = 505.27731
Ridge 10	MSE = 2.77803	MSE = 2.05971	MSE = 515.89174
Lasso 1	MSE = 3.02040	MSE = 2.26517	MSE = 1.87747
Lasso 10	MSE = 3.60263	MSE = 1.80967	MSE = 1.35256
(Hist.)			
(Hist.)			

It appears that ”Ridge 10” performs the best in dataset A, and ”Lasso 10” performed the best for both dataset B and C. The classic ridge regression uses L2 regularization, whereas the Lasso uses the L1 regularization. Hence, the L2 norm will place the outsize penalty on large components of the weight vector, hence, the histogram of L2 (Ridge Regression) would be more evenly distributed across many features. L1 norm leads to a concentrate weights on a small set of features by clearing other weights to zero, which leads to a concentrated and spiky histogram with respect to ridge regression model. The L1 behaviour is also known as feature extraction.

As we can see from the datasets, as the alpha increases, the Lasso becomes more concentrated. In dataset C, the ”Lasso 10” and ”Lasso 1” has some large weights at 10, where classical regression and ridge regression failed to register. For this particular dataset C, Lasso performed the best, hence, we may conclude the dataset C is quite sparse and complex. In the case of sparse data, rigid and classical would be normalized. Hence, we may see the rigid regression with high L2 regularization make worser prediction on the dataset. Whereas, the L1 regularization in Lasso helps to compensate the sparsity and complexity. It is also possible that the ridge regression and linear modelling over-fit the training dataset, while the lasso will reduce this over-fit by feature extraction.

For dataset A, the Ridge with large coefficient performed the best, whereas the lasso with the large coefficient performed the worst. This indicates the dataset A has a significant multi-collinearity. The ridge is capable to shrink the model complexity and multi-collinearity. Whereas, the lasso would reduce the complexity by feature extraction but suffers if the data has a significant multi-collinearity. The lasso would select one of the dominant feature to enhance and attenuates all other features. Hence, the lasso with high regulation factor in this case performed the worst, due to feature losses (or over-simplified).

For dataset B, it is observed that the ridge perform better as the coefficient increases, and the lasso with a large gain perform the best overall. This indicates the dataset has a small degree of complexity.

As per requested on Piazza to think about how we generate the response Y from the true weight distribution $\hat{W}$, and based on the above observation and analysis, we may suspect the response is generated from three different $\hat{W}$ with different distribution model:

The dataset A might generate the weight based on a Gaussian distribution centered at zero mean, resulting a smooth curved histogram centered at $w = 0$. In addition, there are additive noises centered at zero mean with small variance introduced when computing the response. As a result, the ridge performs the best, since there exists multiple features that are correlated at the same time.

The dataset B might be based on a double exponential distribution centered at zero, since the tip of the predicted weight appeared to be quite spiky, and there exists a significant feature resulting a better performance in Lasso.

The dataset C might be based on a double exponential distribution centered at zero as well, but with an additive noise centered at 10, resulting a sparse response Y.