Housing Prices with Ridge Regression

October 19, 2020

Davide Riva, Laurea Magistrale in Data Science and Economics

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

1 INTRODUCTION

It is well known that Ordinary Least Squares estimation of linear regression coefficients may be exceedingly sensitive to some perturbation in training data. More formally, OLS regression is not stable, a desirable property for a learning algorithm. Assuming a square loss function $l(\mathbf{w}) = (\mathbf{w}^\intercal \mathbf{x} - y)^2$, in which case the OLS estimation is the Empirical Risk Minimizer among linear predictors, a way to stabilize it is to add to the empirical risk a regularization term and then to minimize:

$F_\alpha(\mathbf{w}) = \frac{1}{m}\sum_{t=1}^{m}(\mathbf{w}^\intercal \mathbf{x_t} - y_t)^2 + \alpha\|\mathbf{w}\|_2^2 = \frac{1}{m}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \alpha\|\mathbf{w}\|_2^2$

where $\mathbf{X}$ is the $m \times d$ matrix whose rows are samples of the explanatory variables, $\mathbf{y}$ is the vector containing m samples of the target variable and α is a regularization parameter.

This technique, called Ridge regression, aims at reducing variance error and overfitting, but it necessarely introduces a trade-off between stability and accuracy. In particular, training error will get larger and larger as we move away from empirical risk minimization, resulting in a risk of underfitting. However, test error variance, which is used as a proxy for variance error, should shrink.

The scope of the present analysis is to apply such a framework to a real dataset, presented in section 2, in order to check for the improvement in stability, possible underfitting and the response of the algorithm to changes in α. Accuracy on training data will be measured by the R^2 coefficient, whereas the test error (MSE) will be presented explicitly. K-fold cross validation will allow both to weaken the dependency of test error on the specific test set and to empirically measure error variance as $s_K^2 = \frac{1}{K-1}\sum_{k=1}^{K}(MSE_k - \overline{MSE})^2$, where MSE_k is the MSE on the k-th fold and $\overline{MSE}$ is the average MSE on all folds. One drawback of cross validation is that training folds overlap and independency among training errors doesn't hold. As a consequence the estimates of the mean training error and its variance may be biased, and in particular variance may be underestimated. Nevertheless, they will be kept as proxies in order to check for overfitting/underfitting and stability of the model.

Two approaches to the minimization of $F_\alpha(\mathbf{w})$ will be compared: one is the direct computation of the estimated predictor $\hat{\mathbf{w}}$ using the explicit formula $\hat{\mathbf{w}} = (\mathbf{X}^\intercal\mathbf{X} + \alpha m\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y}$, the other involves gradient descent. In the end, I will analyze the behavior of error when applying dimension reduction techniques. Section 3 deals with the details, whereas conclusions are summarized in section 4.

2 DATA

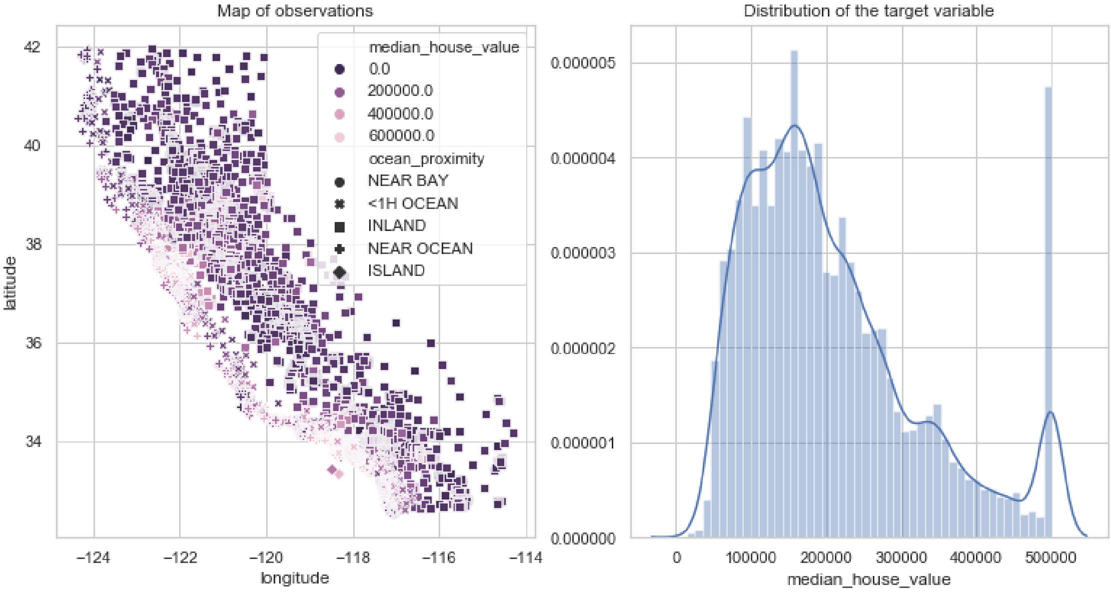
The data were taken from the California housing dataset, which originally contained 20640 observations of 10 features about houses and their neighborhoods. The median house value within a block of houses was chosen as target (dependent) variable for the regression, whereas the explanatory variables are listed below and the first instances shown in the following table:

- longitude
- latitude
- median age of houses within a block
- number of rooms within a block
- number of bedrooms within a block
- population of the block
- number of households in the block
- median income whithin a block
- area (near bay, near ocean, in an hour from the ocean, inland, island)

```
[2]:    longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
     0    -122.23     37.88                41.0        880.0           129.0
     1    -122.22     37.86                21.0       7099.0          1106.0
     2    -122.24     37.85                52.0       1467.0           190.0
     3    -122.25     37.85                52.0       1274.0           235.0
     4    -122.25     37.85                52.0       1627.0           280.0

        population  households  median_income  median_house_value ocean_proximity
     0       322.0       126.0         8.3252            452600.0        NEAR BAY
     1      2401.0      1138.0         8.3014            358500.0        NEAR BAY
     2       496.0       177.0         7.2574            352100.0        NEAR BAY
     3       558.0       219.0         5.6431            341300.0        NEAR BAY
     4       565.0       259.0         3.8462            342200.0        NEAR BAY
```

Plotting a map of instances and the distribution of the target variable, some preliminary observations may be proposed. First of all, houses in the inland generally have lower values than those in other areas, suggesting the ocean proximity categorical variable to be significant, with a negative coefficient for the inland and a positive one for the bay. From this characteristic, also the significance of the longitude feature may be derived, since, other things equal, westernmost houses tend to be more expensive. The same for latitude, with southernmost blocks presenting higher median values.

As for the distribution of the target variable, an odd asymmetry arises. Apart from a little leftward imbalance, a lot of instances unexplainably take a value around 500000. There are two plausible reasons for this behavior: first, data might have been stored improperly, repeating a single entry multiple times; second, values that exceeded the 500000 threshold might have been pushed downward as, for example, in Tillé and Langel's approach for income inequality evaluation[1] or because of privacy issues. Due to this last consideration, as well as to the amount of information they contain (around 5%), such data should not be removed from the analysis. However, their presence contributes much to the imbalance of median house value distribution, and, based on the sensible idea that those instances are part of a totally different market (luxury housing market), I decided to exclude them.



Moreover, 200 observations have a NaN value in the place of the number of bedrooms. Since they represent approximately 1% of the observations, it is reasonable to exclude them altogether, resulting in a dataset of 19448 tuples.

At this point, numeric variables can be standardized so that they have null mean and unitary variance, whereas the categorical one can be converted into 5 dummies (one for each value in its domain). Only 4 out of 5 will be used in the regression, since the fifth one will certainly be collinear with the others. By arbitrary decision, the dummy for *ocean proximity* = *ISLAND* will be excluded. The result of previous operations is exemplified in the following table, again showing the first instances of the dataset.

```
[8]:    longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
     0  -1.329343  1.035347            1.009994    -0.796474       -0.971821
     1  -1.324357  1.026047           -0.588694     2.048480        1.340315
     2  -1.334328  1.021397            1.889272    -0.527944       -0.827460
     3  -1.339313  1.021397            1.889272    -0.616234       -0.720965
```

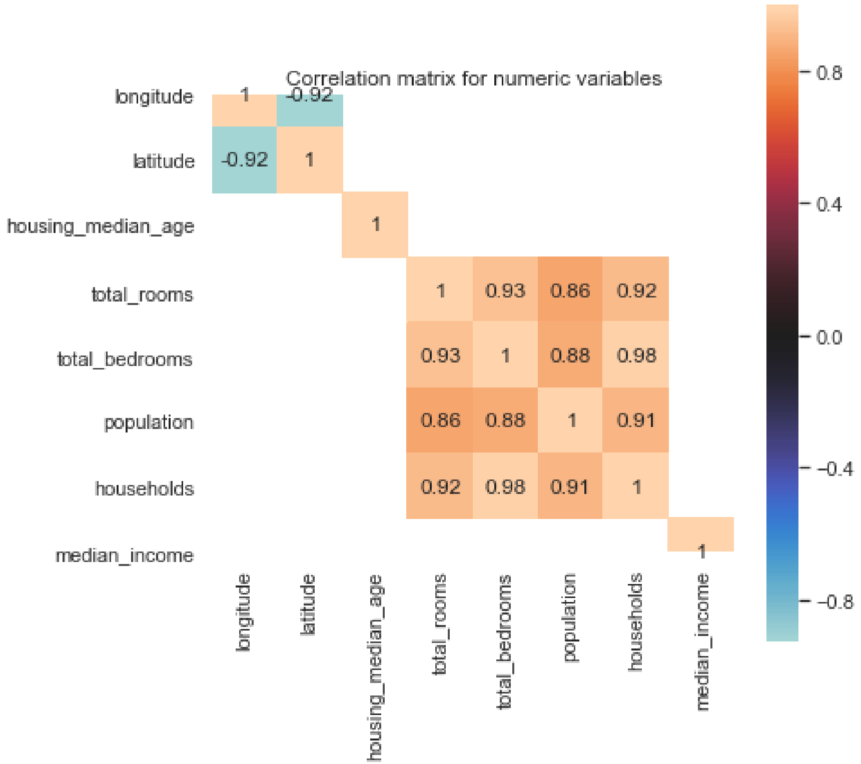
[1] LYON M., LI C., GASTWIRTH J.L., 2016, *The importance of group means for estimating the Lorenz Curve and Gini Index from grouped data*, American Statistician, 70, 25-32

```
4  -1.339313  1.021397            1.889272    -0.454750       -0.614469

   population  households  median_income  median_house_value  <1H OCEAN  \
0   -0.977307   -0.978519       2.961763            2.683772          0
1    0.838681    1.661474       2.946599            1.714624          0
2   -0.825320   -0.845476       2.281460            1.648709          0
3   -0.771163   -0.735911       1.252979            1.537479          0
4   -0.765049   -0.631563       0.108162            1.546748          0

   INLAND  ISLAND  NEAR BAY  NEAR OCEAN
0       0       0         1           0
1       0       0         1           0
2       0       0         1           0
3       0       0         1           0
4       0       0         1           0
```

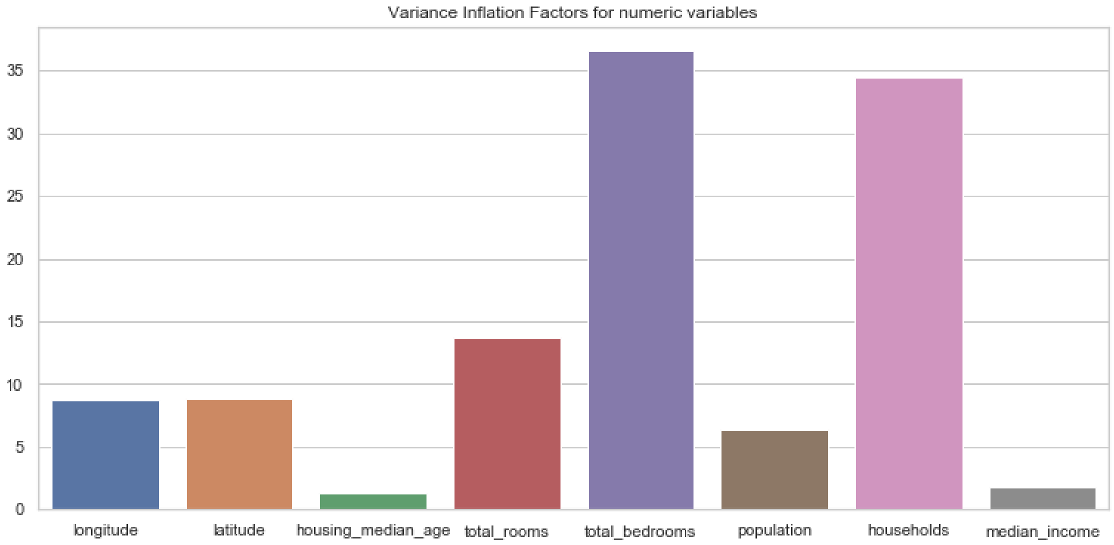
3 ANALYSIS

After pre-processing operations (elimination of NaN and weird values, standardization, definition of dummy variables), it is time to start the analysis.

Multicollinearity To begin with, the presence of multicollinearity issues was checked. The correlation matrix clearly highlights the collinearity between longitude and latitude (due to the peculiar shape of California territory), as well as among the number of rooms, bedrooms, inhabitants and households.

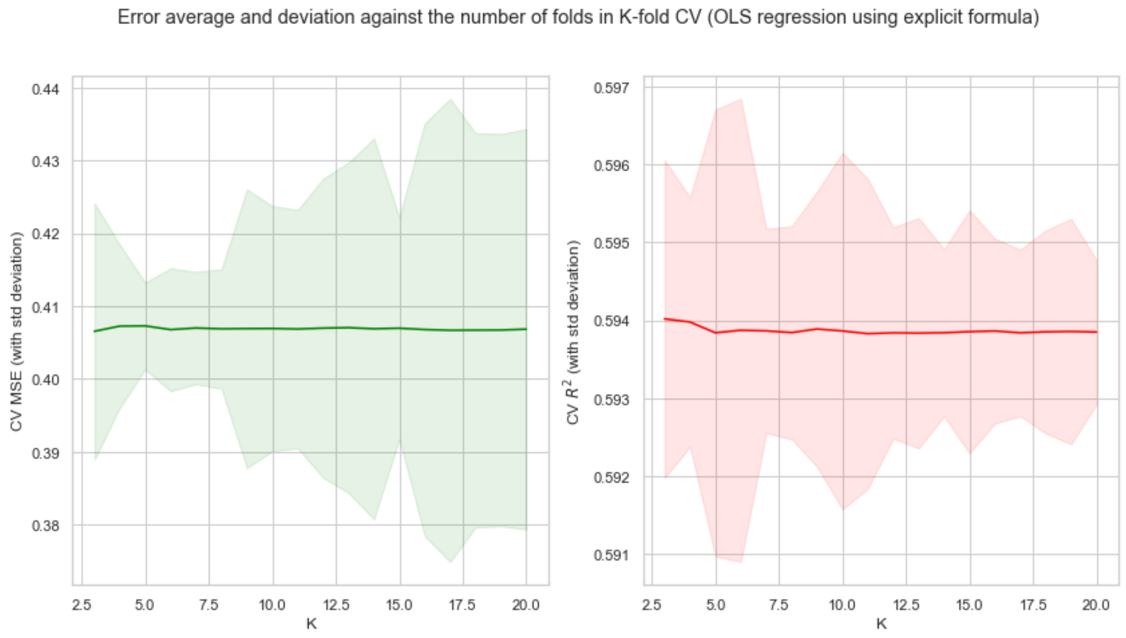


Variance Inflation Factors of numeric features were computed in order to verify these impressions. The following plot shows that the highest VIFs were found for the number of bedrooms and the number of households, which have been subsequently excluded from the regression framework. Among dummies that replaced the categorical variable, the one indicating *ocean proximity* = *ISLAND* was eliminated, as it would surely be collinear with the other 4.



Cross Validation It is now possible to perform the regression and turn to the analysis of the stability-accuracy tradeoff. In this context, cross-validation is useful to define both the error and its variance: as stated in the introduction, here the R^2 coefficient is used to evaluate accuracy on each training fold and the MSE on each test fold; the results are then averaged across folds and their empirical variances are used as proxies to address stability.

One aspect that is worth noticing in this approach is that increasing the number of folds K reduces the however low empirical variance of the R^2, because training folds overlap more, but it also increases MSE empirical variance. The graphs below, in which the explicit formula $\hat{\mathbf{w}} = (\mathbf{X}^\intercal\mathbf{X} + \alpha m\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y}$ was used to compute the predictor, can be considered almost on a par with learning curves, as increasing the number of folds corresponds to enlarging the training set. They show that both test error and training accuracy stay almost constant when K varies, and they confirm that the deviation of the former clearly increases whereas that of the latter tends to decrease.

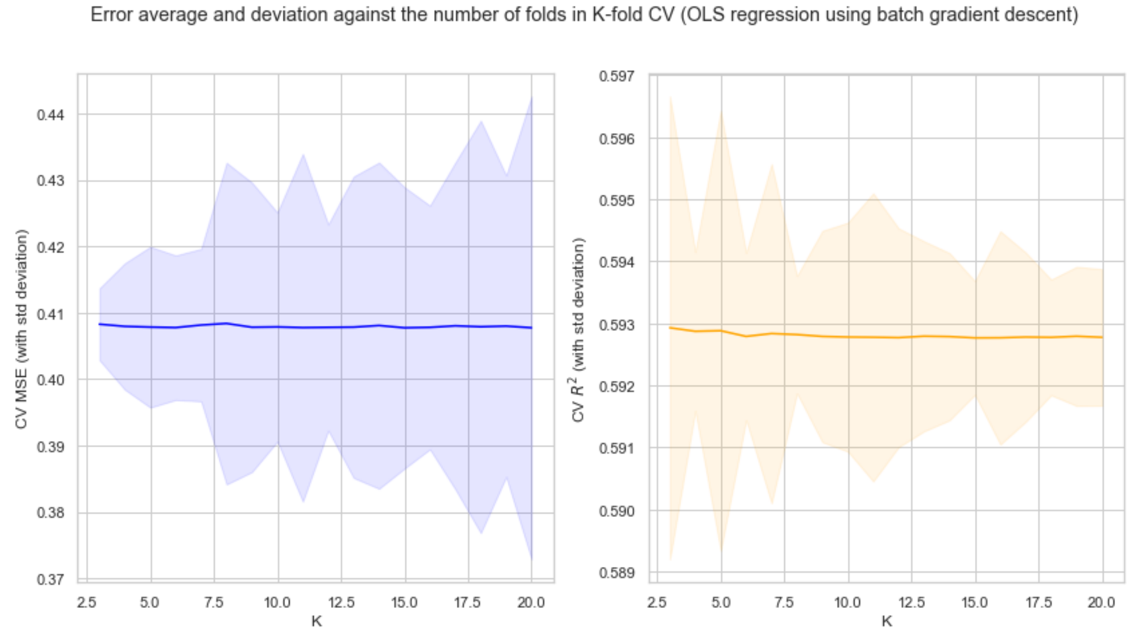

Error average and deviation against the number of folds in K-fold CV (OLS regression using explicit formula)

This behavior is confirmed also when the predictor is computed by batch gradient descent instead of explicit formula. Here, gradient descent was applied with constant learning rate $\eta = 0.25$, as the algorithm diverged in case $\eta = 0.5$ and converged slowly in case $\eta = 0.1$, and precision $\epsilon = 0.01$. This parameter, together with the maximum number of iterations (here set to 1000), shapes the break rule of the algorithm:

- starting from $\mathbf{w_1} = \mathbf{0}$
- while $\|\nabla l_S(\mathbf{w})\| = \|\frac{2}{m}\mathbf{X}^\intercal(\mathbf{Xw} - \mathbf{y})\| > \epsilon$ and $t < 1000$
- update $\mathbf{w_{t+1}} = \mathbf{w_t} - \frac{2\eta}{m}\mathbf{X}^\intercal(\mathbf{Xw_t} - \mathbf{y}) - 2\eta\alpha\mathbf{w_t}$

Again the graphs show constant errors, increasing test error variance and decreasing training error variance. Notice that the confidence intervals are slightly narrower than the ones found by the explicit formula, meaning that the application of gradient descent on different training folds outputs more similar coefficients.

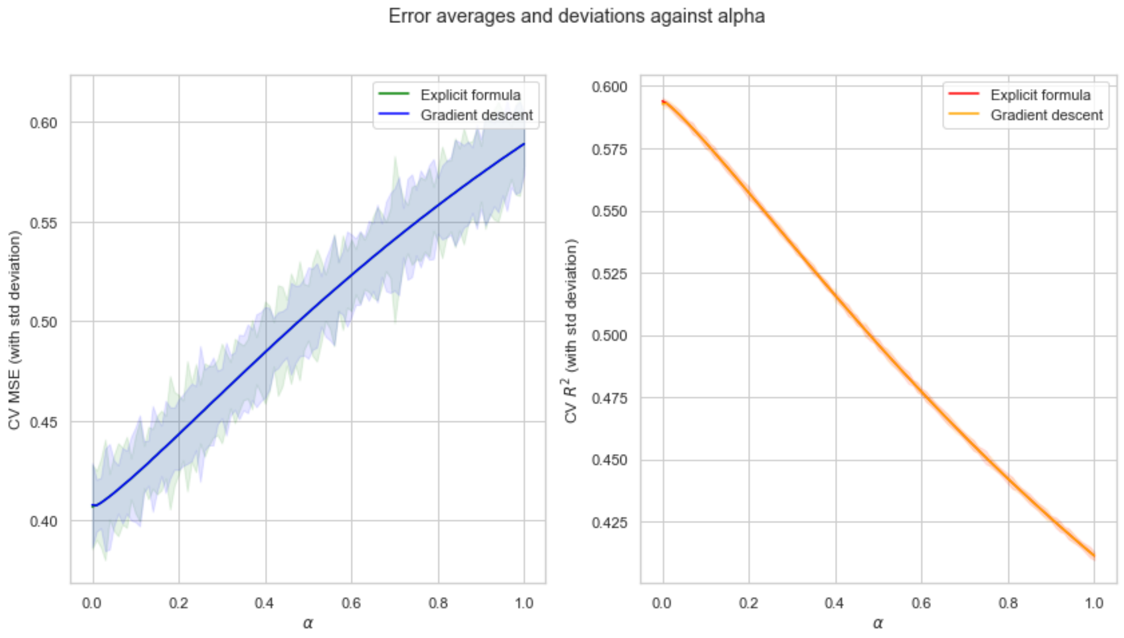
As error empirical variances in the cross validation framework look more or less constant after $K = 10$, such value will be used from now on.



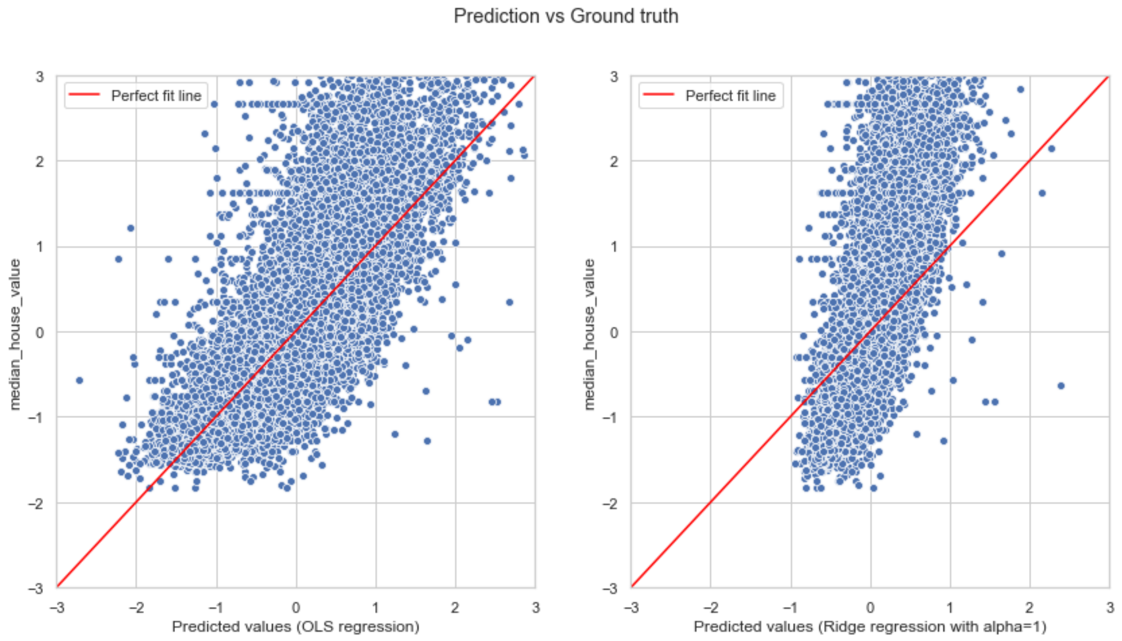
Stability-Accuracy Tradeoff Now that multicollinearity issues have been tackled and the number of folds for cross validation has been chosen, the time has come to introduce the regularization term and examine the behavior of errors and their variances when parameter α changes. Again the estimates are obtained by two different methods (the explicit formula and batch gradient descent) and then compared.

As expected, both results agree on the fact that greater α produce lower accuracy on the training set, with an R^2 dropping from almost 60% to little more than 40%, and that also test error rises from an average of 0.4 to 0.6. Considering that the variance of the target variable has been normalized to 1 and that $R^2 = 1 - \frac{\sum_{t=1}^{m}(y_t - \mathbf{w}^\intercal \mathbf{x_t})^2}{\sum_{t=1}^{m}(y_t - \overline{y})^2} \approx 1 - MSE_{Train}$, the training and test errors follow very similar paths. The model clearly underfits as $\alpha \to 1$ and the monotonic worsening of test error indicates that $\nexists \alpha \in]0;1[$ that minimizes it.

Error standard deviation, which was expected to shrink as α rose, remains in facts almost constant, at least in a range of α from 0 to 1 and except for noisy fluctuations. Also underestimation of training error variance can be seen from the following graph, by comparison with test error variance. One possible interpretation for the absent improvement in error deviation is that the linear model may underfit in any case. Indeed, the R^2 coefficient never reaches 60%. Ridge regression can't bring any improvement from this perspective and other models might be tried.



We may visualize the difference between OLS prediction and Ridge prediction using the following graph, which compares predicted values and empirical values for the target variable. It is evident that Ridge regression tends to shrink predictions towards the mean, and it performs worse than OLS over the whole dataset.



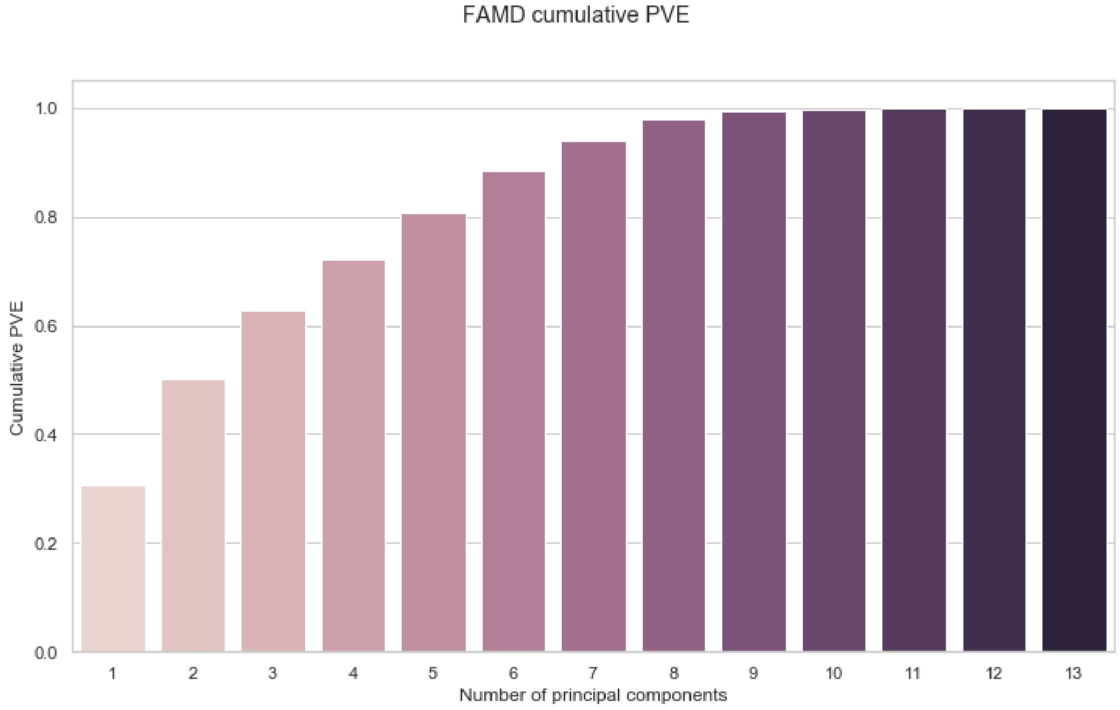
Dimension Reduction As stated in the introduction, another objective of the present analysis is to verify the effects of dimension reduction on the regression performance.

Since the dataset contains both qualitative and quantitative variables, Principal Component Analysis is not appropriate by itself, unless the only categorical feature was considered as supplementary. Leaving this alternative approach for later, I start by applying Factor Analysis of Mixed Data, a particular technique that combines PCA with Multiple Correspondence Analysis.

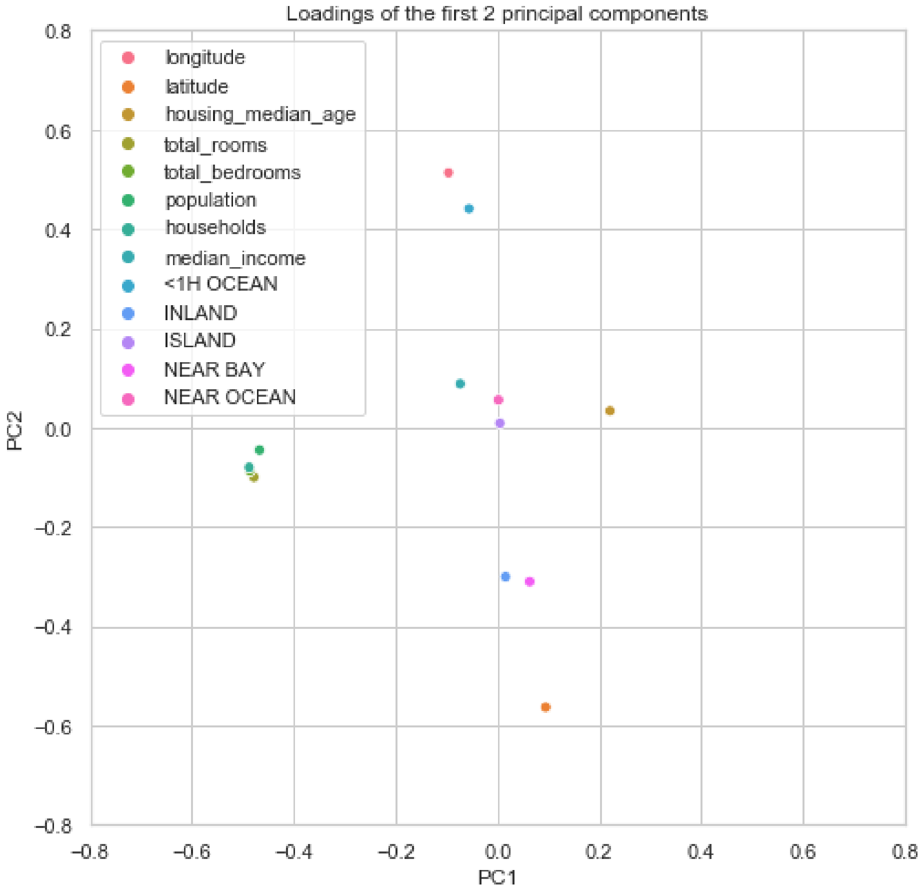
The dataset, which now contains all the initial numeric variables and the 5 dummies, must first be randomly splitted into a training and a test set. Since we have seen in the cross validation framework that error estimates are quite stable over different folds (average dispersion, or variance-to-mean ratio, is 0.1% for test error and far lower for training error), cross validation isn't used in this part of the analysis. Instead, 75% of the observations, amounting to 14642 randomly drawn tuples, is allocated into the training set, whereas the remaining 25% is used for validation.

Since in principle test data are unknown, the principal directions that stem from dimension reduction are computed using only the training data. Let's start from Factor Analysis of Mixed Data. In this technique, dummy variables must first be transformed to fit PCA: more precisely, each of them is divided by the number of its occurrences and is then standardized. PCA is applied to the resulting data frame, leading to the definition of 13 principal components with their associated eigenvalues, which represent the proportion of variance explained by each component $PVE_j = \frac{\lambda_j}{\sum_{i=1}^{13} \lambda_i}$, $j = 1, ..., 13$.

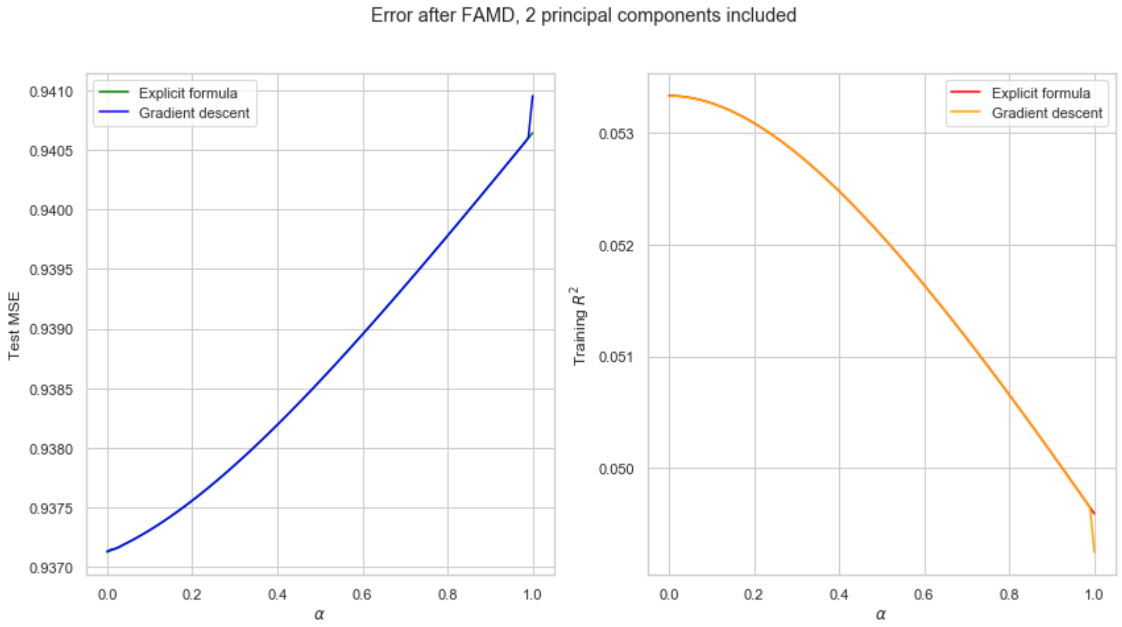
The cumulative screeplot in figure is a simple representation of the growth of PVE when new directions are added. Here it can be noticed that the variance explained by the first 2 principal components amounts to less than 50%, meaning that these directions are generally inappropriate to explain the distribution of data by themselves. At least 5 components would indeed be needed to reach 80%.



Despite being poor in explained variance, the first 2 principal directions are easily interpretable: the first one is negatively correlated with population, the number of rooms, bedrooms and households, but it is positively correlated with housing age, so it is possible to think about a dichotomy between new blocks of large houses (low value of this component) and old blocks of small ones (high value of this component); the second, instead, is almost purely geographical, longitude and local dummies having the highest absolute loadings: the South-East area presents high values in this direction, whereas the North-West presents low ones. Curiously, median income, which was the most relevant variable in previous models, has low correlation with both directions, thus suggesting that dimension reduction won't yield accurate results.

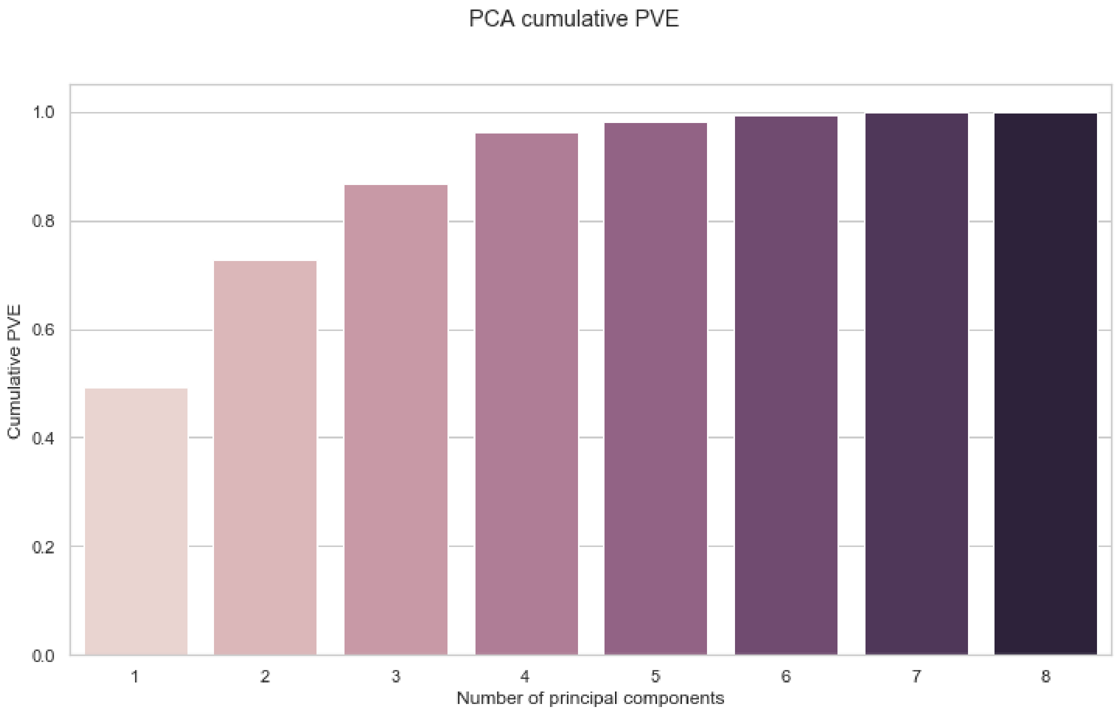


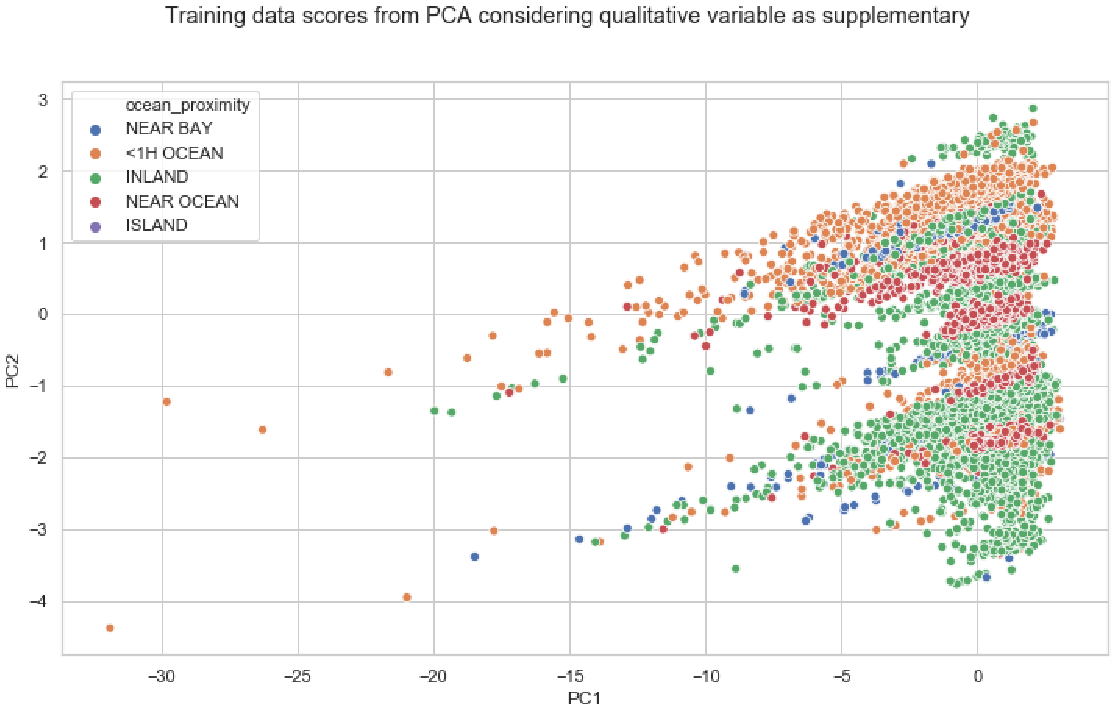
Quite the opposite, a clear worsening is visible. The test MSE rises by 0.44 on average, more than doubling for low α, and the R^2 deteriorates from an average of 51% to almost 5%. Both the computations (by explicit formula and by batch gradient descent) confirm the deterioration, which could be forecasted from the low PVEs of the two principal components in FAMD. The aforementioned absence of the most relevant feature from the two directions and therefore from the regression obviously plays a big role in this finding.



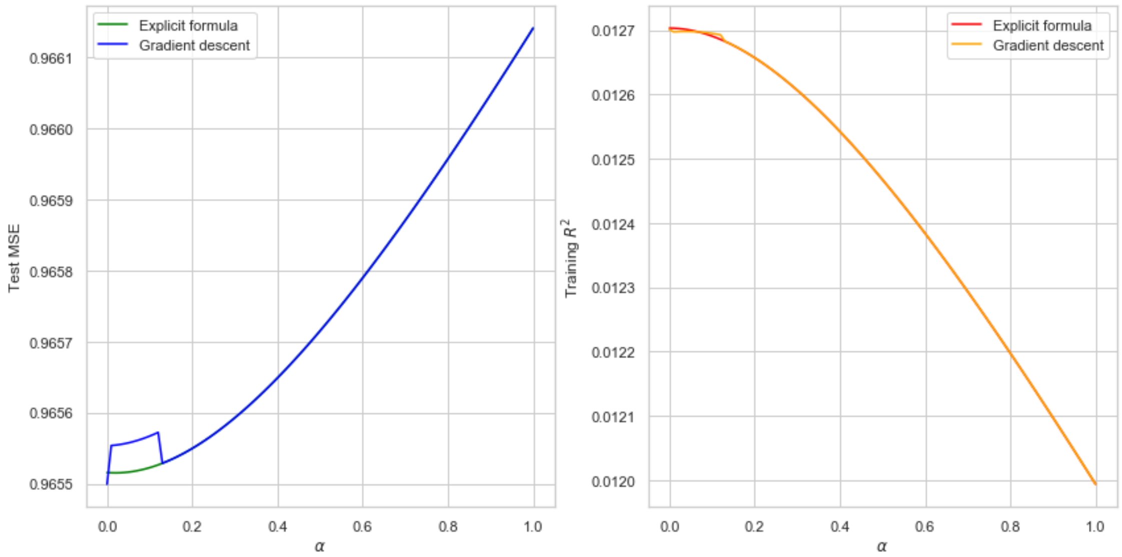
Although dimension reduction yields poor results, an alternative approach is also tried. In fact, the categorical variable can be considered as supplementary, while leaving quantitative features as the only active ones. If the scores obtained throught PCA were clearly clusterable into the 5 geographical areas, then the categories may be explainable in terms of the other variables and PCA may therefore be the correct approach. In particular, principal directions would have higher PVEs, thus improving their representativity of data.

The graphs below show that the cumulative PVE of the 2 principal components indeed rises from less than 50% to more than 70%, although the third component is still needed to reach 80%. Nevertheless, the scatterplot denies the possibility to clearly allocate training points to geographical clusters, so FAMD on all variables may be preferable to PCA applied only to numeric ones.





Implementing Ridge regression on the two principal components, whose interpretation is the same as in FAMD, the situation gets even worse, as shown in the following graphs. This result is proof that excluding the categorical variable may well improve PVEs, or better reduce the overall variance to be explained, but it yields poorer results in the regression framework, thus making FAMD clearly preferable despite lower capability to explain variance of data.



Ridge regression error after PCA without categorical variable, 2 principal components included

4 CONCLUSIONS

On the basis of the previous analysis, the purposes stated in Section 1 can now be checked and the actual results summarized. In particular, the questions of interest were:

- whether Ridge regression would improve stability and how it would affect accuracy;
- how dimension reduction techniques would impact the regression.

The answer to the first one was negative: little to no change in error variance across folds was observed when varying the regularization parameter α. Moreover, both the average MSE on test folds and the average R^2 on training folds tended to get worse as α increased. A plausible explanation was outlined, mainly based on the low training accuracy: although, in principle, regularization should help avoid overfitting, the linear model may underfit, so a broader class of predictors may be needed. Computations were carried on by two different means, either using the closed-form equation or applying batch gradient descent, and both methods agreed on the results.

As for dimension reduction, Factor Analysis of Mixed Data even worsened the results. In fact, when restricting the regression to the first 2 principal components, accuracy fell sharply both on training and test data, probably due to the fact that those dimensions didn't include the variable that had been identified as the major driver for house value, i.e. the median income within the neighborhood. Notice that here a reverse causality issue may arise, but methods to overcome this bias won't be discussed here. Back to dimension reduction, considering the categorical variable as supplementary and applying PCA only to quantitative ones helped improve the Proportion of Variance Explained by the first 2 principal components, but resulted in further worsening of training and test accuracy. Also in this case the two different methods produced very similar output.