Housing Prices

Anna Olena Zhab'yak, Michele Maione

**Summary** –

This paper is trying to answer the question, what is the best model to predict the median house price given the house characteristics aka features.

The classic OLS regression is not enough sometimes to get a good prediction model, for example if features are highly correlated; it has the desired property of being unbiased, however it can have a huge variance. In order to pll down the variance we add some bias with a regularization techniques.

In this paper we implement the Ridge regression and the Lasso regression on a dataset of housing prices in order to predict the median house price given the house characteristics.

# Introduction and description of the problem

This paper starts from a real problem of housing prices in the United States, where it is estimated that household wealth is nearly 50% invested in real estate.

63,5% owner occupied housing rate in July 2019 [[1]](#footnote-1)

The task of predicting the value of a house becomes crucial to the economy as the constructed house price model can influence the growth of the real estate market.

*house price prediction model helps fill up an important information gap and improve the efficiency of the real estate market (Calhoun, 2003)[[2]](#footnote-2).*

An accurate prediction on the house price is important to prospective homeowners, policy makers and other real estate market participants, such as, mortgage lenders and insurers [[3]](#footnote-3)

There are few technical issues with the prediction of the house prices.

The value might be extremely influenced by the value of the neighborhood [[4]](#footnote-4) and so there is a relationship between location and property value; or some exogenous factors can cause an external shock on the overall prices and the prediction becomes wrong.

Moreover working on a large dataset, like the one used in this work, can lead to the so-called multicollinearity of the features which tend to overfit when it comes to implementing certain algorithms to predict the value.

Many works including different techniques have been developed to handle these problems, and they will be introduced later on.

The focus of this paper, however, is on the regularization techniques : the ridge regression, firstly introduced by Hoerl and Kennard (1970), useful tool for improving prediction in regression situations with highly correlated predictors[[5]](#footnote-5);

And Lasso regression

It is a machine learning method that helps to overcome the problem of multicollinearity and therefore to overfitting the data by introducing panalities on the coefficients and the objective function to minimize overall objective function.

The tuning parameter controls the bias - variance trade off, so the choose of it is important.

For choosing the regularization parameter in practice, cross-validation (CV) is widely used.

*and the analysis of the risk minimization given different algorithms will be provided*

this two methods can be used on a dataset with features that represents different formats (example here we have money, people, years, geografical information)

# Most important related works

Many work have been developed to predict the median house value, all of them implementing more or less complex models.

For the hedonic model of price prediction the bundles of characteristics rather than bundles of goods are ranked according to their utility bearing abilities. Attributes (for example, characteristics of a house such as number of bedrooms, number of bathrooms, number of fireplaces, parking facilities, living area and lot size) are implicitly embodied in goods and their observed market prices. The amount or presence of attributes associated with the commodities defines a set of implicit or "hedonic" prices (Rosen, 1974). The marginal implicit values of the attributes are obtained by differentiating the hedonic price function with respect to each attribute (McMillan et al., 1980).

The advantage of the hedonic methods is that they control for the characteristics of properties, thus allowing the analyst to distinguish the impact of changing sample composition from actual property appreciation (Calhoun, 2001)

But

Even though the hedonic price model has been widely recognized, issues such as model specification procedures, multicollinearity, independent variable interactions, heteroscedasticity, non-linearity and outlier data points can seriously hinder the performance of hedonic price model in real estate valuations.[[6]](#footnote-6)

<https://openreview.net/pdf?id=HklRwaEKwB>

Our work is connected to prior works on ridge regression in high-dimensional statistics (Serdobol-skii, 2007

El Karoui (2018)and Dicker (2016) study ridge regression estimators, but focus only on the risk for identity covari-ance. Hastie et al. (2019) study “ridgeless” regression, where the regularization parameter tends tozero.

The influence of the neighbourhood cited before has been studied by Dubin, Robin (2004) and ---

They implemented the kriging algorithm to create an accurate spatial interpolation of house prices in Austin based on the spatial autocorrelation properties of real estate. Implementation of a model to estimate a hedonic regression.

Method of predicting house values is to use data on the characteristics of the area’s housing stock to incorporate for the spatial correlation into the prediction using kriging. [[7]](#footnote-7)

# 3.Notation and relevant definitions (review)

The goal of the ridge regression is to generate a prediction such that the loss function is small for most data points , where is a predictor from the labels set with the regularization term, and data domain; prediction mistakes are a function of the difference.

## Ridge Regression

**constrained minimization** problem

Ridge regression modifies the standard linear regression by introducing a positive constant as regularization parameter.

Starting from the classic linear model we have

Let be the data domain and

. The linear predictor is a linear function of the data points

, where and where wich can be rewritten as

The Regression with squareloss , we want the linear predictor to be as good as possible, for istance we can look at the risk that would be the base optimal

The empirical risk minimization to is

*Where is the vector of predictions and vector of real labels*

*so we can define*

*=*

Let be the design matrix matrix with features and observatios , what we get from data points, that are rows ,

see as the dot product of the two vectors

Applying the ERM we derive

To minimize this convex function

in order to find the minimum value becomes a problem to minimize the eucleadian distance. Using the gradiend descend approach we derive the with respect to w

Least squares solution, we minimize squared error

has to be non singular matrix (Usually it is invertible ) if the data points span r

And second is condition for **general position**

Closed form for erm

Hight bias if the function is non linear or can be nearly linear then my error will be small

If we change exaples, the fit can change significally, it is instable and it leads to variance error. High bias and high variance.

If the inverse in unstable so the matrix non singular

To sability and optimization

This is a linear fit, could find a good solution on the training data, but give a bad estimate on different data. The reason why it occurs it is because of multicollinearity of the prediction vectors (aka not orthogonal [[8]](#footnote-9))

and more in general

with large or small, the risk that the model can overfit[[9]](#footnote-10) the data is high.

The new estimated parameter becomes

,

α is a regularizer parameter

There could be two different situations, for so we are in the prevoious regression

for ,

If alfa goes up, if we minimize over w we do not care aboute the forst term couse the regularizaton will dominate, we want w alfa be small

Shrink the linear regression solution towords to zero, rapresenting by w alfa. W is not stable

W alfa hat has more bias but also less variance

wanted to trade off between bias and

variance to get to that sweet spot of having good predictive performance.

biased estimators of the regression coefficients in the regression model

regularization by introducing some bias ( bias = how well the fit correspond to the true value )

What ridge regression essentially does is tunning the bias variance tradeoff and overcome the problem of overfitting the data.

The solution, to do the fit, for a suitable value of w is given by the optimization problem where the loss function is the objective function and performing the gradient descend it is possible to tune the parmeter alfa to minimize the squared

The gradients then becomes

---minimize this objective function here

solving for w we have

THIS IS THE RIDGE REGRESSION SOLUTION

, is the Lasso regression estimator , where the penalty is multiplied by the absolute value of the coefficient.

Adding the identity fixes the invertibility problem

Two

Always compute inverse, and this is more stable solution

Alfa is the one measuring the stability of the procidure.

The risk of erm that have been stabilised by this alfa.

Risk analisys on the term alfa.

Control the risk of thi by making the solution more stable, more resistant to overfitting.

Alfa goes up, x coefficient goes down and line is flatter .

*In multiple regression it is shown that parameter estimates based on minimum residual sum of squares*

*have a high probability of being unsatisfactory, if not incorrect, if the prediction vectors are not*

*More in generale where D is a large nimber, so we have lots of features, the regression model has a lot of fleibility to explain the data and there is the risk of overfitting*

*Model can rapidly overfit the data, if n small*

*We now introduce, > 0, in the ERM functional:*

**Lasso regression**

Least Absolute Shrinkage and Selection Operator, or simply Lasso, is a method In the same way as ridge regression, Lasso is lasso penalizes the sum of their absolute values (L1 penalty).

The shrinkage of the slope can be zero,

many coefficients are exactly zeroed under lasso, which is never the case in ridge regression.

If alpha increase, we have some parameters go straightway to zero

**Loss function**

*The loss function it is used for* [*parameter estimation*](https://en.wikipedia.org/wiki/Parameter_estimation) *to map an* [*event*](https://en.wikipedia.org/wiki/Event_(probability_theory)) *or values of one or more variables onto a* [*real number*](https://en.wikipedia.org/wiki/Real_number) *intuitively representing some "cost" associated with the event.*

With loss function we denote the measure of how different the prediction of a hypothesis is from the true outcome.

We use a nonnegative loss function to measure the discrepancy between the predicted label and the true label . In the regression task we define the quadratic loss that is the squared distance between and

When then otherwise If and *c* is large then also tend to be large.

y . {\displaystyle y.} The [risk](https://en.wikipedia.org/wiki/Risk_(statistics)) associated with hypothesis h ( x ) {\displaystyle h(x)} is then defined as the [expectation](https://en.wikipedia.org/wiki/Expected_value) of the loss function:

*The risk function , or the expected loss, becomes the mean squared error of the estimate with respect to the probability distribution*

in regression we refer to the estimation of a probability density as the loss function as the eucleadian distance in a function space , ideed:

And the risk function becomes the squared distances between our target variable and predicted values

The L2 Loss [Mean Square Error (MSE)](https://medium.freecodecamp.org/machine-learning-mean-squared-error-regression-line-c7dde9a26b93) is used as loss function to performe this rigde regression.

## Empirical Risk Minimization

The empirical risk minimization defines a family of learning algorithms used to output some predictor minimizing the training error given a non-negative real-valued loss function :

L ( y ^ , y ) {\displaystyle L({\hat {y}},y)} R ( h ) = E [ L ( h ( x ) , y ) ] = ∫ L ( h ( x ) , y ) d P ( x , y ) . {\displaystyle R(h)=\mathbf {E} [L(h(x),y)]=\int L(h(x),y)\,dP(x,y).} The empirical risk minimization principle states that the learning algorithm should choose a hypothesis h ^ {\displaystyle {\hat {h}}} which minimizes the empirical risk:[[10]](#footnote-11)

**Test Error and training error (**generalization error**)**

estimate the predictive power of a predictor, we uuse the test error which allows to performe the validation of the algorithm. It is given by

,

The training error is given to fit the model on training set and performe the model selection and it is

Thebiasis an error term that stems from a mismatch between the model class and the under-lying data distribution, and is typically monotonically non-increasing as a function of the complexity of the model.

Where high complexity models could have very low bias, but high variance.

Whereas low complexity models have high bias, but low variance.

model varianceincreases and bias decreasesmonotonicallywith model com-plexity

BIAS VARIANCE TRADE OFF

The bias is the difference between the true population parameter and the expected estimator:

It measures the accuracy of the estimates.



Graph: the optimal choose for trade-off[[11]](#footnote-12)

In Regularization regression we use the hyper parameter alpha to add a bias , prevent the model from real, it will have better accurancy on the testing set, it means it will not overfit in the training set.

**Cross-validation**

Cross-validation (CV) is one of the techniques used to test the effectiveness of a machine learning models, it is also a re-sampling procedure used to evaluate a model if we have a limited data[[12]](#footnote-13).

In case of limited data, this approach can present a high bias, since there is a substantial loss of data in the training set, but if data is huge and our test sample and train sample has the same distribution then this approach is acceptable.

**K-fold**

K-fold cross-validation is a non-parametric method for evaluating the accuracy of a predictive rule.[[13]](#footnote-14)

it ensures that every observation from the original dataset has the chance of appearing in training and test set

for model selection

Hyperparameters whose value must be determined before the training phase can start

The computed error on the testing part of each fold is

is the dataset divided in subsets , where is the testing part of the

We want to estimate the for a fixed hyperparameter .

And the CV final error is the avarage of the errors

given the choice of two predictors, it repeatedly picks the most accurate of the two. Fix a training\test set

So for tuning the hyperparameter

we look on

where

Estimate risk on all S

This is the algorithm[[14]](#footnote-15)

* for p in 1:P:
* for k in 1:K:
* keep fold k as hold-out data
* use the remaining folds and *λ* = *λp* to estimate β^ridge
* predict hold-out data: ytest,k=Xtest,kβ^ridge
* compute a sum of squared residuals: *S****S****Rk* = ||*y* − *yt****e****st*,*k*||2
* end for k
* average SSR over the folds: SSRp=1K∑k=1KSSRk
* end for p
* choose optimal value: *λo****p****t* = *argminpSSRp*

Moreover the choose of k

he k value must be chosen carefully for your data sample.

A poorly chosen value for k may result in a mis-representative idea of the skill of the model, such as a score with a high variance (that may change a lot based on the data used to fit the model), or a high bias, (such as an overestimate of the skill of the model).

# 4. Proof of a technical result

Before performing the analysis and regression, pre-procession of data is necessary.

The dataset contains 20640 observations and 10 features for each house including the median house value which is the target value that we are trying to predict. Firstly we create the two constants of labels and data domain:

= median\_house\_value

= data frame

The missing values are handled to avoid errors in the execution of the code.There is a categorical feature which represents the distance from the ocean, we assign a numerical value to each category of labels. The dataset is randomized and outliers are dropped.

CORRELATION

Finally, looking at Pearson's correlation matrix a features reduction is performed by taking away the most correlated feature with a treshold of |0.75|

**Cross- validate risk estimate**

Lasso

Cross-validated risk estimate: 0.022512047975181457

Ridge

Cross-validated risk estimate: 0.02580007952062598

[the plot]

as we see the generalized variance is smaller then in training set as this is the k-fold technque that allows to have this results.

Large a ---------------small a

Hight bias low bias

Low varance high variance

Ex w hat = 0 rss

We perform an analysis of different tunning parameters alpha and the relative mean squared error, comparing the Lasso and Ridge (Cholensky)



The best aplha value is 0.2 more or less for ridge regression, then the error increases (>0.022) for values alpha > 0.2

In lasso regression this increase is smoother, and mse is more stable for alpha in between 0.1 and 1.

This is what we could expect if there are feature highly correlated and that are not crucial for the regression.

On the other hand R2 is speculat to the mse and tells u show well the model fits the data



The coefficient magnitude is shown in this graph



Median income has a huge power in the prediction, it means that it drives the values.

# 5. Some critical considerations

Outliers problem

Even if ridge regressioni s modelled in a way to deal with multicollinearity and outliers, it still being sensitive to outliers. In our work there were some outliers that were pushing the fit in a wrong direction.

<https://iopscience.iop.org/article/10.1088/1742-6596/890/1/012150>

<https://iopscience.iop.org/article/10.1088/1742-6596/890/1/012150/pdf>

As the model complexity increases, the models tends to fit even smaller deviations in the training data set. Though this leads to overfitting, lets keep this issue aside for some time and come to our main objective, i.e. the impact on the magnitude of coefficients. This can be analysed by looking at the data frame created above.

It is clearly evident that the **size of coefficients increase exponentially with increase in model complexity**. I hope this gives some intuition into why putting a constraint on the magnitude of coefficients can be a good idea to reduce model complexity.

What does a large coefficient signify? It means that we’re putting a lot of emphasis on that feature, i.e. the particular feature is a good predictor for the outcome. When it becomes too large, the algorithm starts modelling intricate relations to estimate the output and ends up overfitting to the particular training data.

Lasso can set some coefficients to zero, thus performing variable selection, while ridge regression cannot.

# 6. Copyright

We declare that this material, which We now submit for assessment, is entirely our 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 our work. We 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 us or any other person for assessment on this or any other course of study.

This way Lasso performs better in terms of reducing the variance in models with a lot of features (ritondanti)

In contrast Ridge regression performs better in models where a lot of features are important

# Bibliographical references

* Ramsay J. & Huber M. (1987) - The Techniques of Modern Structural Geology. Volume 2, Folds and Fractures. Academic Press, London, 391 pp.
* Ramsay J. & Huber M. (1987) - The Techniques of Modern Structural Geology. Volume 2, Folds and Fractures. Academic Press, London, 391 pp.

Theory:

<https://arxiv.org/pdf/1509.09169;Lecture>

1. United States Census Bureau

   <https://www.census.gov/quickfacts/fact/table/US/HSG445218> [↑](#footnote-ref-1)
2. Calhoun C. A., 2003, “Property Valuation Models and House Price Indexes for The Provinces ofThailand: 1992 –2000”, Housing Finance International, 17(3): 31 –41. [↑](#footnote-ref-2)
3. Visit Limsombunchai,

   <file:///C:/Users/annao/AppData/Local/Temp/2004-9-house%20price%20prediction.pdf> [↑](#footnote-ref-3)
4. Frew J. and B. Wilson, 2000, “Estimation The Connection Between Location and Property Value”, Essay in Honor of James A.Graaskamp, Boston, MA: Kluwer Academic Publishers [↑](#footnote-ref-4)
5. Riccardo A. Maronna [↑](#footnote-ref-5)
6. Visit Limsombunchai

   <file:///C:/Users/annao/AppData/Local/Temp/2004-9-house%20price%20prediction.pdf> [↑](#footnote-ref-6)
7. Dubin, Robin. (1998). Predicting House Prices Using Multiple Listings Data. The Journal of Real Estate Finance and Economics. 17. 35-59. 10.1023/A:1007751112669. [↑](#footnote-ref-7)
8. https://www.math.arizona.edu/~hzhang/math574m/Read/RidgeRegressionBiasedEstimationForNonorthogonalProblems.pdf [↑](#footnote-ref-9)
9. Overfitting : ….. [↑](#footnote-ref-10)
10. V. Vapnik (1992). [<http://papers.nips.cc/paper/506-principles-of-risk-minimization-for-learning-theory.pdf> *Principles of Risk Minimization* for Learning Theory.*]* [↑](#footnote-ref-11)
11. Source : researchgate.net [↑](#footnote-ref-12)
12. <https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f> [↑](#footnote-ref-13)
13. <https://arxiv.org/pdf/1909.11696.pdf> [↑](#footnote-ref-14)
14. <https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net> [↑](#footnote-ref-15)