ON THE PREDICTION ACCURACIES OF THREE MOST KNOWN REGULARIZERS : RIDGE REGRESSION, THE LASSO ESTIMATE, AND ELASTIC NET REGULARIZATION METHODS

Adel Aloraini

Computer Science department, Qassim University, Saudi Arabia.

ABSTRACT

The work in this paper shows intensive empirical experiments using 13 datasets to understand the regularization effectiveness of ridge regression, the lasso estimate, and elastic net regularization methods. the study offers a deep understanding of how the datasets affect the goodness of the prediction accuracy of each regularization method for a given problem given the diver- sity in the datasets used. the results have shown that datasets play crucial rules on the performance of the regularization method and that the predication accuracy depends heavily on the nature of the sampled datasets.

KEYWORDS

ridge regression, regularization, the lasso estimate, elastic net .

1. INTRODUCTION

Penalization over regressive parameters have taken much attention in the literature recently. This is clearly due to the optimality such methods can provide in the model se- lections over non-penalized regressive parameters(such as in linear regression). The way that penalized regressive methods shrink the parameters associated with features in the model, is key in providing better predictions when model selection is required among models in the search space. The most well known penalized regressive methods can be seen in ridge regression[1], lasso estimate[2], and recently elastic net regularization method[3]. In the ridge regression, the model selection is not much severed in which the regressive param- eters are penalized towards ”zeros” which usually introduces light sparsity to the model selection and includes all the chosen subset features from the search space. However, lasso estimate introduces much severity and usually the penalization term (λ) discards irrelevant features from the chosen model[2]. However, the shortcoming from lasso estimate is when the search space includes features which are highly correlated. The lasso estimate in a highly correlated feature space tends to choose one among the correlated features and ignores the others which,if were chosen, might lead to a better prediction accuracy. In addition, the lasso estimate selects at most (n) features when the number of features (p) is greater than the number of samples(n) which indicates that the number of features is bounded by the number of samples[3]. Elastic net regularization method on the other hand, can handle limitations encountered by the lasso estimate in terms of : (1) choosing the group of features that best give better prediction, and (2) giving unbounded behavior by the number of samples in ($p >> n$) situation.

DOI: 10.5121/ijaia.2017.8603

Hence, given the formentioned interesting diversities between regularisation methods, we show in this work empirical exper- iments to step further in analyzing model selection behavior in terms of prediction accuracy for ridge regression, the lasso estimate, and elastic net regularization methods.

To curry out model selection, we used Bayesian information criteria(BIC) and cross validation score functions for all fitted models in the search space by using a set of different (λ) values as well as different (α) values for elastic net regularization method.

In the next section, we detail the related work then we proceed to detail the methodology, the results, and the conclusion

2. RELATED WORK

Ridge regression, the lasso estimate, , and lately elastic net regularization methods have been extensively used for model selections and feature reductions in machine learning literature and applications. In [4] ridge regression has been applied in a combination approach between homomorhpic encryption and Yao garbled circuits which outperformed using homomorphic encryption or Yao circuits only. Ridge regression also has shown interesting results when multi-colinearity exist in model selection and associated parameters specially when ridge re- gression is combined with Variance Inflation Factors(VIF) [5], [6]. However, ridge regression is often used when features in the model selection are all important to be included in resultant classifiers or models. However, when the search space encounters many irrelevant features to the problem under modeling, then lasso estimate can investigate and return sparser models that contain the most important features. this is because lasso estimate shrinks many associated parameters towards zero that tend to be less relevant. In [7] the paper has experimented with the lasso estimate and its variants and the results have shown that when the lasso estimate is relaxed with filtering methods, the prediction is improved. Also, the lasso estimate in [8] has been applied for visual object tracking and the results have shown promising performance for computer vision field. Moreover, for network modeling, the work in [9] has addressed how useful the lasso estimate is to estimate psy-chopathological networks specially that estimated parameters fast growing comparing to the data samples. However, the work has concerned that the lasso estimate can yield a sparse models that can capture interesting results usinge exponential growth parameters in the search space under investigation. However, when there exists a group of features naturally correlated and co-work with each other such as between genes in cellular systems , often the lasso estimate tends to choose a member of the group and ignores the others [3] as well as the bounders imposed by the number of chosen features which is subject to the sample size [3]. The aforementioned shortcomings from the lasso estimate have already been addressed and solved by the elastic net regularization methods which deals with the correlated features either to be all in-or-out of the selected model. Elastic net regularization method also shows a more reliable feature selection in the $p >> n$ datasets and due to the scalability elastic net regularization method provides in the model selection and in the estimated parameters, it has drawn much attention recently. Elastic net regularization method in [10] was able to determine a reliable performance with high accuracy in assessing the uncertainties on node voltage and

determine the influential factors along with calculating their voltage influence parameters. In the study by [11] a hybrid probabilistic prediction methodology based on elastic net regularization method with probabilistic Bayesian belief network has been applied to predict the on-show probability of the patient to the clinic using demographics, socioeconomic status as well as available appointment information history which provide a notable comparable predictive results among many approaches in the literature.

In our current study we aim to provide an intensive compar- ison between the lasso estimate, ridge regression, and elastic net regularization methods in order to further analyze the prediction accuracy of these methods. We used 13 datasets that are different in sample size and the systems that have been sampled from to apply different diversities to the behavioral analysis for each penalized method as well as the score functions used to do model selection.

In the next section we go in details for the methodology and the datasets used in the study.

3. DATASETS

The datasets used in the study have been obtained from different systems. We used gene expression datasets from microarray experiments that differ in sample and feature size. The datasets are annotated by different signaling pathways : (1) cell-cycle signaling pathway which has 5 samples and 98 genes that considered to be hard-to-learn-from as the feature space is far exceeds the sample space. (2) MAPK signaling pathway has the same gene size as cell-cycle signaling pathway but with 100 samples. We also used two different microarray datasets sampled from prostate cancer gene expression signal- ing pathways : (1) JAKSTAT signaling pathway with 86 genes across 13 samples, and (2) JAKSTAT1 signaling pathway that has 35 genes v.s. 62 samples. A much harder-to-learn-from datasets we used come from breast-cancer tissues : (1) the breast cancer dataset1 contains 209 genes vs. 14 samples, and the breast cancer dataset2 contains 209 samples vs. 10 samples. In order to allow diversity to the methods in the experiments, we used 7 datasets from equities market that allow for big sample size comparing to the feature space that are banks in the market. The banks in the equities market have been monitored in 7 intervals as follows : (1) equities-dataset1 has 11 features vs. 80 sample size,(2) equities-dataset2 has 11 features vs. 59 sample size, (3) equities-dataset3 has 11 features vs. 39 sample size , (4) equities-dataset4 has 11 features vs. 19 sample size,(5) equities-dataset5 has 11 features vs. 7 sample size, (6) equities-dataset6 has 11 features vs. 250 sample size, and (7) equities-dataset7 has 11 features vs. 250 sample size.

4. METHODOLOGY

All 13 datasets described above have been injected to the three penalized regressive methods : ridge regression, the lasso estimate, and elastic net regularization methods. In ridge regression, the lasso estimate, and elastic net regularization methods the set of (λ) values have been computed by the func- tion (glmnet) in R for 100 values [12], [13]. Thus, each dataset generates a search space of 100 models corresponding to 100 λ values. The generated models then have been scored by two scoring functions, namely cross validation (Algorithm 1) and Bayesian information criterion (BIC) (Algorithm 2) . When using BIC score function, all models were generated by the corresponding (λ) values and then BIC scores all these models and returns from the search space the model with the smallest BIC score. For cross validation we set (nfolds=3) as we have small

sample size in some datasets used in the experiments. For the elastic net regularization method, we experimented with 8 different (α) values that were between [0.1, 0.9] since α=1.0 corresponds to the lasso penalty and when α=0.0 it gives ridge regression penalty. After defining the necessary parameters, (λ, α, and nf olds)), we executed the experiments over the 13 datasets. Algorithm 1, and Algorithm 2 describe the framework of applying cross validation score function, and BIC score function respectively on ridge regression, the lasso estimate, and elastic net regularization methods.

In the next sections we explain the component of Algorithm 1 that gives how cross validation was used from within the pre- defined (λ,α) to choose the most optimal model in the search space of each dataset used in the experiment, and Algorithm 2 explains the same but for BIC score function.

A. Algorithm 1

Algorithm (1) starts in step(1) by defining the vector of (α) used in the experiments that ranges between [0.0,1.0], where (0.0) penalizes for the lasso estimate and (1.0) penalizes for ridge regression. Then, in step(2) the algorithm iterates on each value of α.vector for the dataset under investigation in step(3). The algorithm iterates on the length of feature space to experiment with each feature(step(6)) to find out the most optimal subset of features from the remaining features in the dataset(step(5)). To do that, first ; in step(7), cross validation score function with the pre-defined((nf olds))is used to score all possible models generated from the set of (λ) values(100 values between [0-1]) for the possible subset of features from step(5). Second, the algorithm in step(8) returns the best optimal value of (λ) that corresponds to the smallest error from cross validation score function which in turns is used in step(9) in order to return the best fit model in the search space for the current(αi). Finally , in step(10), and step(11) the best fit model is used as a predictive model to estimate the goodness-of-fit for the chosen optimal (λ).

Algorithm 1 This algorithm describes the framework of applying cross validation score function with the three penalised regularizers

1: $\alpha.vector = i[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$
2: **for** $i = 1$ **to** $length(\alpha.vector)$ **do**
3: $dataset = READ(dataset)$
4: **for** $z = 1$ **to** $length(FeatureSpace(dataset))$ **do**
5: $X = dataset[, -z]$
6: $Y = dataset[, z]$
7: $cv.score = cv.glmnet(X, Y, nfolds = 3, \lambda = [0, 1 : 100])$
8: $optimal_\lambda = min(cv.score_\lambda)$
9: $fitted.model = glmnet(X, Y, \alpha = \alpha_i, \lambda = optimal_\lambda)$
10: $predictive.model = predict(fitted.model, X)$
11: $predictive.error = mean((Y - predictions)^2)$
12: **end for**
13: **end for**

B. Algorithm(2)

Algorithm (2) starts in step(1) by defining the vector of (α) used in the experiments that ranges between [0.0-1.0] , where (0.0) penalizes for the lasso estimate and (1.0) penalizes for ridge regression. Then, in step (2) the algorithm iterates on each value of α. vector for the dataset under investigation in step(3).vector for the dataset under investigation in step(3). After that the sample size parameter is determined in step (4) in order to be used in BIC score function later. The algorithm in step (5) iterates on the length of feature space to experiment with each feature (step(7)) in order to find out the most optimal subset of features from the remaining features in
To do that, first ; in step(8) all pre- defined (λ) are used to fit all the possible models in the search space for a particular (α) . Then, the algorithm in step (9) extracts the exact (λ) values used to generate all models in the search space. After that, in step(10) the algorithm iterates on each value of (λ. vector) to generate a model(step(11)),predict on the fitted model(step(12)), calculate the prediction error(step(13)), and then calculate the number of features found in the model(step(14)) which can be determined by the nonzero parameters(s) in the model. After that, the algorithm calculates the BIC scoring function as in step (15). All BIC score functions for all models are stored in a vector (step(16)) in order to be used to return the best BIC score function as in step(18). After that the best BIC score function is used to return the best model as in step (19), then finally the prediction accuracy for the chosen model is calculated in step (20), and step (21).

Algorithm 2 This algorithm describes the framework of applying BIC score function with the three penalised regularizers.

1: $\alpha.vector = i[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$
2: **for** $i = 1$ **to** $length(\alpha.vector)$ **do**
3: $dataset = READ(dataset)$
4: $size = length(sample(dataset))$
5: **for** $z = 1$ **to** $length(FeatureSpace(dataset))$ **do**
6: $X = dataset[, -z]$
7: $Y = dataset[, z]$
8: $fitted.model = glmnet(X, Y, \alpha = \alpha_i, \lambda = [0, 1 : 100])$
9: $\lambda.vector = \lambda(fitted.model)$
10: **for** $l = 1$ **to** $length(\lambda.vector)$ **do**
11: $fitted.model = glmnet(X, Y, \alpha = \alpha_i, \lambda = \lambda.vector[l])$
12: $predictive.model = predict(fitted.model, X)$
13: $predictive.error = mean((Y - predictive.model)^2)$
14: $num.features = length(which(\beta(fitted.model)! = 0.0))$
15: $BIC = size * log(predictive.error/size) + num.features * log(size)$
16: $BIC.vector < -append(BIC.vector, BIC)$
17: **end for**
18: $min.BIC = min(BIC.vector)$
19: $fitted.model = return(fitted.model(min.BIC))$
20: $predictive.model = predict(fitted.model, X)$
21: $predictive.error = mean((Y - predictive.model)^2)$
22: **end for**
23: **end for**

5. RESULTS AND DISCUSSION

The algorithms described in the previous section, have been applied to the aforementioned 13 datasets. Table .I shows the results of applying the methodology described in Algorithm 1, and Table .II shows the results of applying the methodology described in Algorithm 2. The range of (α) values from 0.0-1.0 were used in order to experiment with the lasso estimate (when α=0.0), ridge regression (when alpha=1.0), and elastic net regularization methods. When datasets were used to experiment with the methodology in Algorithm 1, cross validation score function did not significantly show a better (α) over another in terms of prediction accuracy except for cell-cycle dataset when α= 0.7,0.8,0.9,1.0 , and for MAPK dataset when α= 0.6,0.7,0.8,0.9,1.0 in which the prediction accuracy considered to be the worst among other α values. Hence the lasso estimate worked better than ridge regression and elastic net regularization methods for these particular datasets. Similarly, Table. II shows the results of applying the methodology described in Algorithm 2. When datasets were used to experiment with the methodology in Algorithm 2 BIC score function has shown similar prediction accuracies for the datasets across different values of (α) comparing to cross validation score function in Algorithm 1 except for 6 datasets in which BIC score function has given a better prediction accuracy comparing to cross validation. These datasets are : equities-dataset5 when α={0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, cell- cycle dataset when α={0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, MAPK dataset when α= 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 , JAKSTAT1 dataset when α= 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 , breast cancer dataset1 when α= 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 , and breast cancer dataset2 when α= 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 as the prediction accuracy was almost (~ 0.0). As a result from Table-I, the lasso estimate, ridge regression, and elastic net regularization methods tend to work similarly expect for cell-cycle , and MAPK datasets in which ridge regression outperformed the lasso estimate and elastic net regulariza- tion methods. In Table-II, the lasso estimate and elastic net regulariztion methods have shown a better results than ridge regression in ~47% of the datasets but still work similarly in the other datasets. On looking thoroughly for the score functions used in the comparison, the average of prediction accuracies for all datasets across all different values of α were considered and it can be seen that BIC score function outperformed cross validation score function as shown in Figure(fig:prediction).

Final Prediction Accuracy for all datasets vs. alpha values 0.0

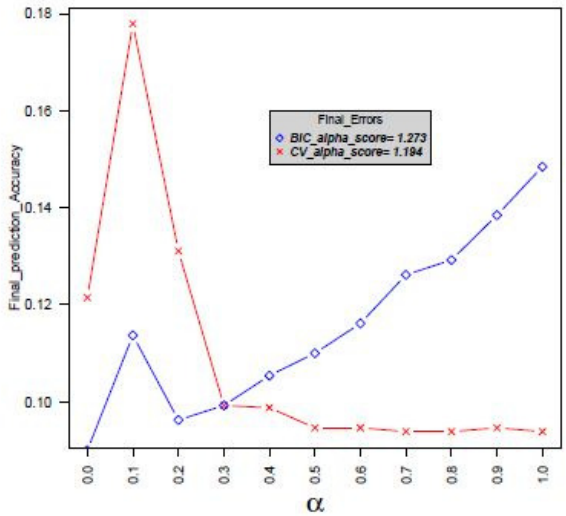

Fig. 1. This figure shows the average of prediction accuracies for all datasets across all different values of . The figurte shows that BIC score functionoutperformed cross validation score function.

α	eq-ds1	eq-ds2	eq-ds3	eq-ds4	eq-ds5	eq-ds6	eq-ds7	cel-cy	MAPK	JAK	JAK1	breast	breast2	AvgErr
0.0	0.14	0.11	0.07	0.01	0.009	0.41	0.41	~0.0	~0.0	0.01	0.001	~0.0	~ 0.0	0.09
0.1	0.41	0.11	0.07	0.02	0.01	0.41	0.41	0.005	0.009	0.01	0.01	0.002	0.002	0.11
0.2	0.14	0.11	0.07	0.02	0.01	0.41	0.41	0.01	0.02	0.02	0.02	0.006	0.005	0.09
0.3	0.14	0.11	0.07	0.01	0.01	0.41	0.41	0.02	0.04	0.02	0.03	0.01	0.01	0.1
0.4	0.14	0.11	0.08	0.02	0.02	0.41	0.42	0.04	0.06	0.02	0.03	0.01	0.01	0.1
0.5	0.14	0.11	0.07	0.01	0.02	0.41	0.41	0.06	0.09	0.02	0.05	0.02	0.02	0.11
0.6	0.14	0.11	0.08	0.02	0.02	0.41	0.41	0.08	0.13	0.02	0.05	0.02	0.02	0.11
0.7	0.14	0.11	0.08	0.02	0.02	0.42	0.41	0.10	0.16	0.03	0.09	0.03	0.03	0.12
0.8	0.14	0.11	0.08	0.02	0.03	0.41	0.41	0.12	0.19	0.02	0.08	0.03	0.04	0.12
0.9	0.14	0.11	0.08	0.02	0.04	0.41	0.41	0.15	0.23	0.03	0.09	0.04	0.05	0.13
1.0	0.14	0.12	0.08	0.02	0.03	0.41	0.42	0.18	0.29	0.03	0.12	0.04	0.05	0.14

TABLE I. THIS TABLE SHOWS THE RESULTS OF APPLYING THE METHODOLOGY DESCRIBED IN **ALGORITHM 1**

α	eq-ds1	eq-ds2	eq-ds3	eq-ds4	eq-ds5	eq-ds6	eq-ds7	cel-cy	MAPK	JAK	JAK1	breast	breast2	AvgErr
0.0	0.17	0.12	0.08	0.01	0.05	0.5	0.5	0.02	0.03	0.01	0.06	0.02	0.009,0.1	10.12
0.1	0.14	0.11	0.07	0.01	0.003	0.41	0.41	0.26	0.37	0.05	0.2	0.16	0.12	0.17
0.2	0.14	0.12	0.07	0.01	0.002	0.41	0.41	0.18	0.19	0.05	0.002	0.05	0.07	0.13
0.3	0.14	0.11	0.07	0.01	0.01	0.41	0.41	0.02	0.04	0.02	0.03	0.01	0.01	0.1
0.4	0.14	0.11	0.08	0.01	0.002	0.42	0.42	0.03	0.01	0.05	~ 0.0	0.008	0.004	0.09
0.5	0.15	0.11	0.08	0.01	~ 0.0	0.42	0.42	~ 0.0	~ 0.0	0.04	~ 0.0	~ 0.0	~ 0.0	0.09
0.6	0.15	0.11	0.08	0.01	~ 0.0	0.42	0.42	~ 0.0	~ 0.0	0.04	~ 0.0	~ 0.0	~ 0.0	0.09
0.7	0.15	0.11	0.08	0.01	~ 0.0	0.42	0.42	~ 0.0	~ 0.0	0.03	~ 0.0	~ 0.0	~ 0.0	0.09
0.8	0.15	0.11	0.08	0.01	~ 0.0	0.42	0.42	~ 0.0	~ 0.0	0.03	~ 0.0	~ 0.0	~ 0.0	0.09
0.9	0.15	0.12	0.08	0.01	0.0	0.42	0.42	0.0	0.0	0.03	0.0	0.0	0.0,	0.09
1.0	0.15	0.11	0.08	0.01	~ 0.0	0.42	0.42	~ 0.0	~ 0.0	0.03	~ 0.0	~ 0.0	~ 0.0	0.09

TABLE II. THIS TABLE SHOWS THE RESULTS OF APPLYING THE METHODOLOGY DESCRIBED IN **ALGORITHM 2**

6. CONCLUSION

The study in this paper focused on how the ridge regression, the lasso estimate, and elastic net regularization methods behave in terms of prediction accuracy when wrapped up with BIC, and cross validation score functions in 13 different datasets that are different in dimensionality. The results clearly show that the performance of a single regularizer is subject to the dataset under investigation which makes the prediction accuracy differ accordingly. The results also show that the lasso estimate and elastic net regularization methods perform better compared with ridge regression and this is a justification that ridge regression includes more irrelevant features than the lasso estimate and elastic net in the chosen model which decreases accuracy in the prediction.

REFERENCES.

[1] A.N. Tikhonov And V.Y. Arsenin.Solutions Of Ill-Posed Problems.Wiley, New York, 1977.

[2] Robert J. Tibshirani. Regression Shrinkage And Selection Via The Lasso.Journal Of The Royal Statistical Society, Series B, 58(1):267–288, 1996.

[3] H. Zou And T. Hastie. Regularization And Variable Selection Via The Elastic Net. Journal Of The Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2003.

[4] Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis, Marc Joye, Dan Boneh, And Nina Taft. Privacy-Preserving Ridge Regression On Hundreds Of Millions Of Records. In Ieee Symposium On Security And Privacy, Pages 334–348. Ieee Computer Society, 2013.

[5] Bonsang Koo And Byungjin Shin. Using Ridge Regression To Improve The Accuracy And Interpretation Of The Hedonic Pricing Model : Focusing On Apartments In Guro-Gu, Seoul. In Ieee Symposium On Security And Privacy, Volume 16, Pages 77–85. Korean Institute Of Construction Engineering And Management, 2015.

[6] C.B. Garca, J. Garca, M.M. Lpez Martn, And R. Salmern. Collinearity: Revisiting The Variance Inflation Factor In Ridge Regression. Volume 42, Pages 648–661, 2015.

[7] Adel Aloraini. Ensemble Feature Selection Methods For A Better Regu- Larization Of The Lasso Estimate In P>>N Gene Expression Datasets. In Proceedings Of The 12th Conference In Machine Learning And Applica- Tions, Pages 122–126, 2013.

[8] Qiao Liu, Xiao Ma, Weihua Ou, And Quan Zhou. Visual Object Tracking With Online Sample Selection Via Lasso Regularization. Signal, Image And Video Processing, 11(5):881–888, 2017.

[9] Sacha Epskamp, Joost Kruis, And Maarten Marsman. Estimating Sychopathological Networks: Be Careful What You Wish For. Volume 12, 2017.

[10] Pengwei Chen, Shun Tao, Xiangning Xiao, And Lu Li. Uncertainty Level Of Voltage In Distribution Network: An Analysis Model With Elastic Net And Application In Storage Configuration. In Ieee Transactions On Smart Grid, 2016.

[11] Kazim Topuz, Hasmet Uner, Asil Oztekin, And Mehmet Bayram Yildirim. Predicting Pediatric Clinic No-Shows: A Decision Analytic Framework Using Elastic Net And Bayesian Belief Network. Annals Of Operations Research, 2017.

[12] Jerome Friedman, Trevor Hastie, And Robert Tibshirani. Regularization Paths For Generalized Linear Models Via Coordinate Descent. Journal Of Statistical Software, 33(1):1–22, 2010.

[13] Noah Simon, Jerome Friedman, Trevor Hastie, And Rob Tibshirani. Reg- Ularization Paths For Cox's Proportional Hazards Model Via Coordinate Descent. Journal Of Statistical Software, 39(5):1–13, 2011.