Measurement of Performance using Regression and Statistical Analysis of Big Data

Anuj Puranik[1], Gagan Sharma[2] and Rajesh Sharma[3]
Department of Computer Science and Engineering, Sri Satya Sai College of Engineering
RKDF University, Bhopal, India
[1]anuj87in@gmail.com,[2]gagansharma.cs@gmail.com, [3]rajeshsharma.ercs@gmail.com

Abstract: Statistical and regression analysis are significant methods for data analysis, now it can be applied to analyze big data also. This paper presents a performance matrix of Gaussian and random big data analysis using statistical and regression techniques. The complexity of big data processing increases as volume of the data increases, so ordinary tools and methods are not appropriate for this analysis. The modeling scale for big data reduces the volume, then statistical and regression analysis can be applied to the mapping modeled big data. Simulation result based on ridge regression analysis on Gaussian and random big data presents a new approach toward big data analysis.

Keywords: Big Data, Statistics, Regression, Data Analysis, MapReduce

1. Introduction

With the development of the term big data, the elaboration of modern digital world has boosted to new technology. The large volume of complex and growing data generated from many distinct sources has led to the era of Big Data. Companies depend on this massive data to take intelligence de- cisions as well as to gain a powerful competitive advantage. The new world is also depends mostly on the big data that uses in business, hospital industry and government organizations. Hence the big data is extremely useful in taking intelligent business decisions in many fields. In large volumes of data so much useful values are hidden and that can be generated only through the careful analysis.

We have come into the world of big data in any case, more and more big data issues are deriving in fields, such as scientific research, and international economics, public ad- ministration and so on. Discovering techniques to big data became a current eventuality and challenge. In big data applications, it is usual for the attribute to work in the same manner with response or explanatory variable. Various tech- niques of analysis was developed in view of such problems, like logistic regression,and k-nearest neighbor method so on. In any way, some papers express an important type of big data issue in which the response variable works as the real numeric type and the explanatory variable is the attribute type.

1.1 Technical View of Big Data

Big data means extremely a big data, it is a collection of large datasets that can't be handled using conventional computing techniques. Big data is not only a data, rather it has turned into an entire subject, which includes different tools, techniques and systems. Big data includes the data delivered by various devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

Black Box Data: It is a part of helicopter, airplanes, and jets, and so forth. It catches voices of the flight group, recordings of receivers and headphones, and the execution information of the aircraft.

Social Media Data: Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.

Power Grid Data: The power grid data holds data utilized by a particular node with respect to a base station.

Search Engine Data: Search engines fetch lots of data from different databases.

Stock Exchange Data : The stock exchange information holds data about the 'purchase' and 'offer' decisions made on an offer of various organizations made by the clients.

Transport Data: Transport information includes model, capacity, distance and availability of a vehicle.

Thus Big Data combines huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

- **Structured data** : Relational data.
- **Semi Structured data** : XML data.
- **Unstructured data** : Word, PDF, Text.

2. Regression Analysis

Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. Regression analysis helps to understand how the typical value of the de- pendent variable changes when any one of the independent variables is varied.

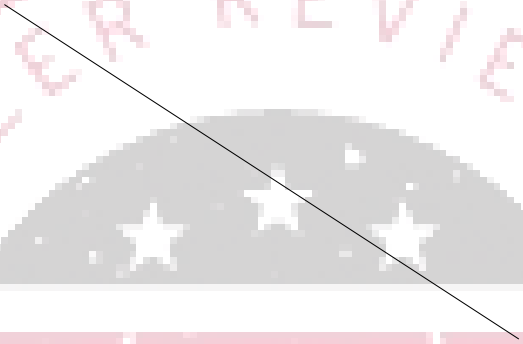


Figure 1: Linear Regression

It is a mathematical function that predicts sales, profits, temperature etc. can be predicted utilizing regressiontech- niques. Basically Regression analysis is to find the value of variable when the value of another variable is given. This technique is used for forecasting and prediction. It is an important tool for modeling and analyzing Data

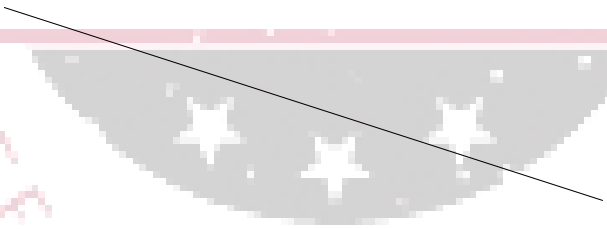


Figure 2: Component of Linear Regression

In ordinary least squares, the regression coefficients are estimated represented in formula

$$\hat{\mathbf{B}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\underline{\mathbf{Y}}$$

Note that fact since then the variables are standardized, $\mathbf{X}^t\mathbf{X} = \mathbf{R}$, where $\mathbf{R}$ is the correlation matrix of independent variables. These approximations are unbiased so that the expected value of the estimates are the population values. That is,

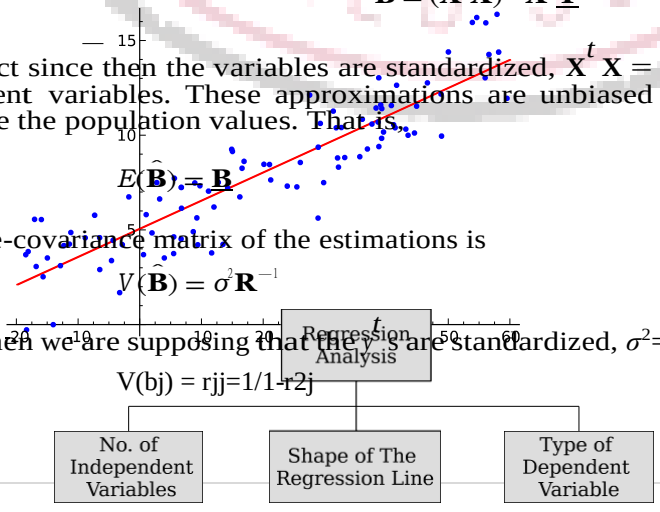
$$E(\hat{\mathbf{B}}) = \underline{\mathbf{B}}$$

The variance-covariance matrix of the estimations is

$$V(\hat{\mathbf{B}}) = \sigma^2\mathbf{R}^{-1}$$

and since then we are supposing that the y_i's are standardized, σ^2=1. From the above, we find that

V(bj) = rjj=1/1-r2j



Where R2 is the R-squared value obtained from regression Xj on the other independent variables. In that case, this variance is the VIF. We look that as the R-squared in the divisor gets closer and more like one, the difference will get larger and larger. The rule of thumb cut-off magnitude for VIF is 20. Solving in reverse, this converts into an R-squared value of 0.19. Consequently, whenever the R-squared value between one independent variable and the rest is more prominent than or equivalent to 0.19, you should face multicollinearity. Now, ridge regression advances by including a small value, k, to the diagonal elements for the correlation matrix. This is the place ridge regression gets its name since the diagonal of ones in the correlation matrix might be consideration of as a ridge.

That is shown as,

$$\tilde{\mathbf{B}} = (\mathbf{R} + k\mathbf{I})^{-1}\mathbf{X}^{t}\underline{\mathbf{Y}} \qquad -$$

k is a positive amount less than one (mostly less than 0.3). The measure of bias in this estimator is predicted by

$$E\ \tilde{\mathbf{B}} - \underline{\mathbf{B}} = (\mathbf{X}^{t}\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^{t}\mathbf{X} - \dot{\mathbf{I}} - \underline{\mathbf{B}}$$

and the covariance matrix is given by

$$V\ \tilde{\mathbf{B}} = \dot{\mathbf{X}}^{t}\mathbf{X} + k^{\Sigma-1}\mathbf{I}\ \mathbf{X}^{t}\mathbf{X}\ \dot{\mathbf{X}}^{t}\mathbf{X} + \underline{k}\mathbf{I}^{\Sigma-1}$$

It can be shown that there consists a value of k for which the mean squared error (the variance plus the bias squared) of the ridge estimator is less than that of the least squares estimator [2]. Unfortunately, the appropriate value of k de- pends on knowing the true [3] regression coefficients (which are being estimated) and an analytic solution has not been discovered ensures the optimality of the ridge solution.

Ridge regression is used as a part of highly correlated multi-independent factor related dependent variables. It is worked to reduce the effect of all factor on the any other. It is a process for examining multiple regression data that suffer from multicollinearity. When multicollinearity happens, least squares estimates are unbiased, but their differences are large so there may be far to the true value. From computing a degree of bias to the regression approximations, ridge regression undermine the standard errors [1].

3. Literature Review

Saritha *et al.* [20] described the various tracks in the process of statistical modeling. Sub datasets are taken from population, unlike traditional statistical analysis and Various types of regression techniques like simple, multiple and multivariate regression models are applied on these sub datasets. Simulations have been done and the results are plotted and compared, an optimal regression model has then found out. The results are tested and compared using the data of bike renting to the public. In this research, a new method is pro- posed to serially partition big data in to sub sets instead of samples to overcome the limitation of big data analytics.

The process of predictive modeling can be divided into various steps as depicted in Figure 3. The first task in a data analytics is to define a measurable and quantifiable goal. Data collection and management step is used to identifying the data we need exploring it and conditioning it to be suitable for analysis. Managing missing values fundamentally do two things.

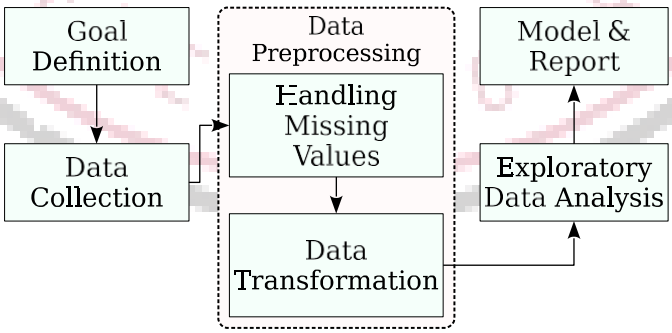


Figure 3: Process Steps of Predictive Modeling

Drop the tuple with missing values or to convert the missing values to a meaningful value. The pur- pose of data transformation is to make data easier to model and easier to understand. Some of the data transformation techniques are converting continuous variable to discrete, normalization and

standardization. Exploratory data analysis is the initial step in predictive modeling. It covers the summary of the data numerically and graphically. Finally the results are interpreted.

Traditional statistical analysis mainly focused on the sampling method for inference of population. As sampling of big data may not give a correct prediction, they propose to partition the Big Data in to sub datasets which is tiny in size and closed to sample. These sub datasets are analyzed with regression technique separately and the results are integrated to estimate the prediction. The architecture for the predictive modeling using this partitioning technique is given in Figure 4.

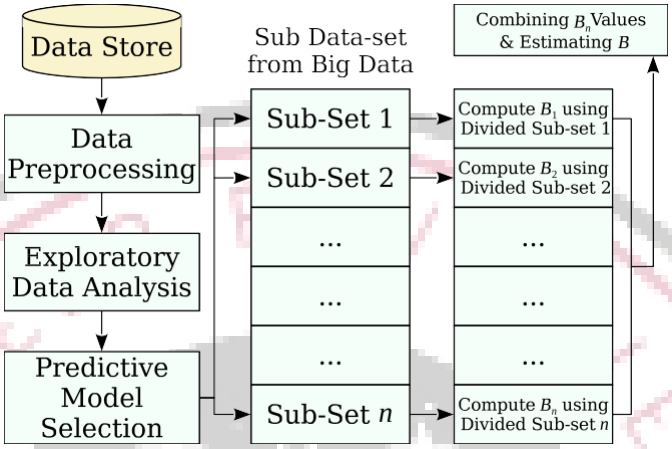


Figure 4: Big Data Partitioning Architecture

In this architecture the whole data set is preprocessed and EDA is performed to find the independent and dependent variable. If the dependent variable is quantitative then we can use Linear Regression technique as prediction model. In regression techniques we have to take a model for the rela- tionship among a response variable (often referred to as the dependent variable) and one or more explanatory variables (often referred to as the independent variables) [1]. The linear regression model is represented by the equation,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (1)$$

Where Y is the response variable and X1, X2, . . . , Xn are independent variables. β0, β1, . . . , βn are regression parameters.

Zhang et al. [21] proposed a ridge regression approach to big data. Ridge regression is important and has been extensively used in applications. The classical ridge regression approach focuses on small or moderate data. It assumes that the entire data set can be loaded to the memory of a personal computer. However, if the data set is large, then it cannot be loaded to memory, which means that the classical approach cannot be used. To solve the problem, they pro- posed new methods and algorithms, where the entire data set is only scanned once. The goal of scanning data is to compute a matrix of sufficient statistics [22], which is not large. Once the matrix of sufficient statistics is derived, all of the rest computations can be completely carried out without the need of the original data set. Therefore, our numerical algorithms are efficient. Uses of ridge regression are still well known in many research areas. Cases include pat- tern and face recognition Xue et al. [23] in 2009, genetics, and machine learning. Previous methods and algorithms for ridge regression are for the most part produced for small or, then again moderate data. They can't be utilized to ex- amine big data as a result of the presence of memory and computational efficiency barriers.

Ridge regression has turned out to be popular and all Around acknowledged since it was first proposed by Hoerl *et. al* [24] in 1970. At the point when multicollinearity hap- pens, despite the fact that the standard thing least squares estimators are as yet unbiased, their variances are absolutely inflated. By including a small degree of bias to the least squares estimators, ridge regression can significantly reduce their standard errors and in this manner [25] increment the levels of significance. The goal of the present article is to propose new methods and algorithms for ridge regression which can overcome these challenges. Ridge regression is one of the most common techniques to enhance the power.

Obenchain *et al.* [26] proposed a method for testing general linear hypotheses in multiple regression models. it is shown that non-stochastically shrunken ridge estimators yield the same central ratios and *t* statistics as does the least squares estimator. Thus although ridge regression does produce biased point estimates which deviate from the least squares solution, ridge techniques do not generally yield "new" normal theory statistical inferences: in particular, ridging does not necessarily produce "shifted" confidence regions. A concept, the ASSOCIATED PROBABILITY of a ridge estimate, is defined using the usual, hyper ellipsoidal confidence region centered at the least squares estimator. To conquer these difficulties, new thinking in statistics and computer science is needed Fan *et al.* [27]2014. Traditional algorithms perform well only in moderate data. If the whole data set is loaded to memory of

a computer, then standard algorithms can be applied. Examples include the computation of the rank statistics, the order statistics, and the standardization [28]. However, these algorithms are infeasible in big data.

4. Proposed Approach

The size to determine whether a data set is large is rela- tive to the available computing resources. It is suggested that a data set is considered large if it exceeds 20% of ran- dom access memory (RAM) on a single computer and mas- sive if it exceeds 50% (Emerson and Kane 2012). Since the size of big data may be much higher than the storage vol- ume of a single computer, the input data may be stored in multiple disks and the computations have to be distributed across hundreds of thousand of computers so that the job can be finished in reasonable amount of time. Therefore for big data, methods and algorithms based on a single processor and a cluster of pro- cessors are both important. Properties of methods in the first case are more important as methods of the second case are often developed based on methods of the first case.

Our goal is to propose an approach to the computation of the exact ridge regression parameters for big data. Our approach contains statistical methods and numerical algo- rithms. We assume that the entire data set cannot be loaded to the memory but it can be saved to the hard disk of a computer. We focus on methods and algorithms based on a single processor. It is critical in selecting important explana- tory variables in the ridge regression approach. Proposed techniques include Linear Rigression, which is the simplest form of Rigression. It models a randam variable, Y called a response variable, as a linear function of another variable X which is called as Predictor Variable. Thus the equation becomes according to linear Regression is :

$$y = a + bX$$

Where the variance of Y is assumed to be constant, a and b are regression coefficients which specifies the Y intercept and slope of line. The coefficients can be resolved with the technique for Least Squares, which helps in minimization of the data between the actual data and the estimated line.

5. Result Analysis

We conducted a simulation study to evaluate the performance of our approach. All of our computations were carried out by a third generation Intel core-i7 2.8 GHz processor with 8 GB DDR3 memory.

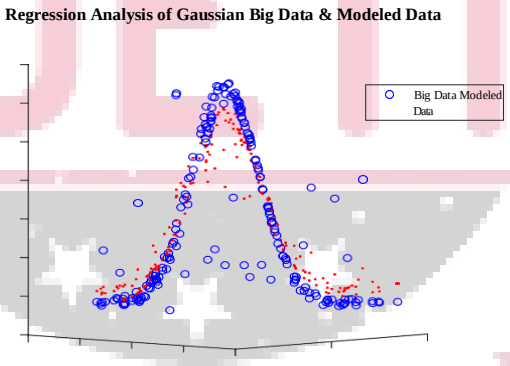


Figure 5. Proposed Method

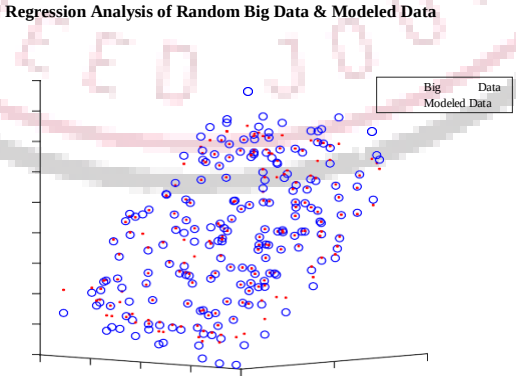


Figure 5. Proposed Method

Figure 5 represents the MATLAB simulation result as comparative Regression analysis of Gaussian big data and modeled data taken as model of big data. Similarly, Figure 6 represents the MATLAB

simulation result as Regression analysis of random big data with comparison of modeled data derived from big data.

The simulation involves one response to be ridge regressive. Data points for the covariates, denoted axis such as (x_1, x_2, y), are drawn from a multivariate normal distribution. The points of x_1 is between (2 to -2), and points of x_2 is also (2 to -2) and points of y axis is between (-1.5 to 2). Table 1 presents a comparative result between actual data

Table 1: Comparison of Result: (Actual Data *vs.* Predicted Data)

Actual Data			Predicted Data		
x_1 [−2, 2]	x_2 [−2, 2]	y[−1.5, 2]	x_1 [−2, 2]	x_2 [−2, 2]	y[−1.5, 2]
1.12	1.3	-0.5	1.16	1.29	-0.56
1.25	1.5	-0.7	1.3	1.47	-0.75
1.5	1.8	-1	1.56	1.87	-0.95
1.3	1.6	-1.1	1.27	1.55	-0.99
-0.85	-1.2	0	-0.90	-1.18	0.2
-1.6	-1.7	0.5	-1.7	-1.66	0.46
-1.8	-1.87	0.9	1.77	-1.8	1
-1.75	-1.9	1	-1.8	-2	1.1
-1.9	-2	1.3	-1.86	-1.59	1.39

5.1 Result Validation

For validation of the comparative result, we used Cross- Conformal Prediction (CCP). The Cross-Conformal Pre- diction (CCP) partitions the training set into K subsets S_1, S_2, . . . , S_K also, computes the resistance scores of the cases in each subset S_k and of $(x_{l+1}, \tilde{y})$ for each possible label. For Big Data, we assume the value of $S_k = 1.5(TB)$ in Table

Such that S_k is the subset and the range og $k = 1$, assume the value of l is $500(MB)$, than we calculate $p(\tilde{y})$. To determine the Modeled data we assume the value of $S_k = 47410(MB)$, and the value of l is $15(MB)$, than the value of $p(\tilde{y})$ is approximate nearest to the Big Data. Table 2 presents the Validation of result using cross-conformal prediction (CCP), it validates that the analysis of actual big data and modeled data have approximately same characteristics after big data modeled reduction.

Table 2: Validation using Cross-Conformal Prediction (CCP)

Parameters	Big Data	Modeled Data
$\sum_{k=1}^{K} S_k$	1582750 (mb)	47410 (mb)
l	500 (mb)	15 (mb)
$p(\tilde{y})$	3165.5	3160.666

6. Conclusion and Future Work

In this research, an approach was proposed to overcome the computing load in big data analysis since mostly Ridge Regression as statistical methods were focused on small sample Gaussian data. In the experimental results, the regression parameters estimated by the big data were not different to the parameters by sub data sets. This research contributes to avoid the computing problem in many fields for big data analysis. In this paper, we have proposed an approach to overcome the computing load in big data analysis since mostly statistical methods were focused on small sample data. And in Big data analysis, we should analyze whole data which are recognized as population in statistics, and this data collection is so large. In our experimental results, we know that the regression parameters estimated by the big data were not different to the parameters by sub data sets.

In our research, we find that statistical quantities are also important. More diverse methods of big data sampling are needed in the future works. According to the findings in our research, we believe that the conclusion can be generally true in many classical statistical approaches and this should be an interesting research topic in the future.

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[2] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, pp. 407–499, 04.

[3] J. S. Vitter, "Algorithms and data structures for external memory.," *Found. Trends Theor. Comput. Sci.*, vol. 2, no. 4, pp. 305–474, 2006.

[4] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," 01 2004.

[5] M. R. Thakare, S. W. Mohod, , and A. N. Thakare, "Article: Various data mining techniques for big data" IJCA Proceedings on International Conference on Quality Upgradation in Engineering Science and Technology.

[6] M. Enea, "Fitting linear models and generalized linear mod- els with large data sets in r,"

[7] J. Polo, D. Carrera, Y. Becerra, M. Steinder, and I. Whal- ley, "Performance-driven task co-scheduling for mapreduce environments," in *2010 IEEE Network Operations and Man- agement Symposium - NOMS 2010*, pp. 373–380, April 2010.

[8] A. Fernández, S. del Río, V. López, A. Bawakid, M. J. del Jesus, J. M. Benítez, and F. Herrera, "Big data with cloud computing: an insight on the computing environment, mapreduce, and programming frameworks," *Wiley Interdis- ciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 5.

[9] D. Miner and A. Shook, *MapReduce Design Patterns: Build- ing Effective Algorithms and Analytics for Hadoop and Other Systems*. O'Reilly Media, Inc., 1st ed., 2012

[10] S.Guha, R Hafen, J Rounds, J Xia, J li, B Xi, and W.S Cleveland "Large Complex Data : divide and recombine with rhipe, Stat , Vol 1, no. 1

[11] P. Ma and X. Sun, "Leveraging for big data regression," *Wiley Interdisciplinary Reviews: Computational Statistics, Vol 7 no. 1*

[12] N. Lin and R. Xi, "Aggregated estimating equation estima- tion.," *Stat. Interface*, vol. 4, no. 1, pp. 73–83, 2011

[13] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of various decision tree algorithms for classifica -tion in data mining," *International Journal of Computer Applications, Vol 163, pp 15-19, Apr 2017*

[14] R. V. Hogg and A. T. Craig, *Introduction to mathematical statistics.(5"" edition)*. Upper Saddle River, New Jersey: Prentice Hall 1995

[15] S. Jun and S.-J. L.-B. Ryu, "A divided regression analysis for big data," *International Journal of Software Engineering and its Applications. Vol 9, no 5, 2015*

[16] L. R. Nair and S. D. Shetty, "Article: Research in big data and analytics: An overview," *International Journal of Com- puter Applications*, vol. 108, pp. 19–23, December 2014. Full text available.

[17] "Privacy-preserved big data analysis based on asymmetric imputation kernels and multiside similarities," *Future Gen- eration Computer Systems*, vol. 78, no. Part 2, pp. 859 –866, 2018

[18] A. Gupta, R. Pandey, and K. Verma, "Article: Analysing distributed big data through hadoop map reduce," *Interna- tional Journal of Computer Applications*, vol. 129, pp. 26– 31, November 2015. Published by Foundation of Computer Science(FCS), NY, USA

[19] W. Q. Meeker and Y. Hong, "Reliability meets big data: Opportunities and challeng es," *Quality Engineering*, vol. 26, no. 1, pp. 102–116,2014

[20] K. Saritha and S. Abraham, "Prediction with partitioning: Big data analytics using regression techniques," in *2017 In- ternational Conference on Networks Advances in Computa- tional Technologies(NetACT), pp 208-214, July 2017*

[21] T. Zhang and B. Yang, "An exact approach to ridge regres- sion for big data," *Computational Statistics*, vol. 32, pp. 909–928, Sep 2017

[22] "Local online kernel ridge regression for forecasting of urban travel times," *Transportation Research Part C: Emerging Technologies*, vol. 46, no. Supplement C, pp. 151-178, 2014

[23] H. Xue, Y. Zhu, and S. Chen, "Local ridge regression for face recognition," *Neurocomputing*, vol. 72, no. 4, pp. 1342– 1346, 2009. Brain Inspired Cognitive Systems (BICS 2006) / Interplay between Natural and Artificial Computation

[24] A. E. Hoerl and R. W. Kennard, "Ridge regression: Bi- ased estimation for nonorthogonal problems," *Technomet- rics*, vol. 12, no. 1, pp. 55–67, 1970.

[25] H. Zhan and S. Xu, "Adaptive ridge regression for rare vari- ant detection," *PLOS ONE*, vol. 7, 08 2012.

[26] R. L. Obenchain, "Classical f-tests and confidence regions for ridge regression," *Technometrics*, vol. 19, no. 4, pp. 429– 439, 1977