mathematics

MDPI

Review

Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review

Jireh Yi-Le Chan [1,*,†], Steven Mun Hong Leow [1,†], Khean Thye Bea [1], Wai Khuen Cheng [2], Seuk Wai Phoong [3], Zeng-Wei Hong [4] and Yen-Lin Chen [5,*]

[1] Faculty of Business and Finance, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia; steven.utar@1utar.my (S.M.H.L.); beakheanthye@1utar.my (K.T.B.)
[2] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia; chengwk@utar.edu.my
[3] Department of Management, Faculty of Business and Economics, Universiti Malaya, Kuala Lumpur 50603, Malaysia; phoongsw@um.edu.my
[4] Department of Information Engineering and Computer Science, Feng Chia University, Taichung 407102, Taiwan; zwhong@fcu.edu.tw
[5] Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei 106344, Taiwan
* Correspondence: jirehchan@utar.edu.my (J.Y.-L.C.); ylchen@mail.ntut.edu.tw (Y.-L.C.)
† These authors contributed equally to this work.

Abstract: Technologies have driven big data collection across many fields, such as genomics and business intelligence. This results in a significant increase in variables and data points (observations) collected and stored. Although this presents opportunities to better model the relationship between predictors and the response variables, this also causes serious problems during data analysis, one of which is the multicollinearity problem. The two main approaches used to mitigate multicollinearity are variable selection methods and modified estimator methods. However, variable selection methods may negate efforts to collect more data as new data may eventually be dropped from modeling, while recent studies suggest that optimization approaches via machine learning handle data with multicollinearity better than statistical estimators. Therefore, this study details the chronological developments to mitigate the effects of multicollinearity and up-to-date recommendations to better mitigate multicollinearity.

Keywords: multicollinearity; variable selection methods; optimization approaches; neural network; machine learning

MSC: 62M10

Citation: Chan, J.Y.-L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.-W.; Chen, Y.-L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics* **2022**, *10*, 1283. https://doi.org/10.3390/math10081283

Academic Editors: Wanquan Liu, Xianchao Xiu and Xuefang Li

Received: 7 March 2022
Accepted: 10 April 2022
Published: 12 April 2022

1. Introduction

Multicollinearity is a phenomenon that can occur when running a multiple regression model. In this age of big data, multicollinearity can also be present in the field of artificial intelligence and machine learning. There is a lack of understanding of the different methods for mitigating the effects of multicollinearity among people in domains outside of statistics [1]. This paper will discuss the development of the methods chronologically and compile the latest methods.

Forecasting in finance deals with a high number of variables, such as macroeconomic data, microeconomic data, earnings reports, and technical indicators. Multicollinearity is a common problem in finance as the dependencies between variables can vary over time and change due to economic events. Past literature tried to remove collinear data to reduce the effects of multicollinearity. This is done through stepwise regression that eventually arrives at a model with a low root mean square error (RMSE). The computational difficulty of this has led to many selection criteria to be developed to choose models. A breakthrough

method to solve multicollinearity came in the form of ridge regression. Instead of selecting variables, all the variables are used. The method modifies the estimator by adding a penalty term to the ordinary least square (OLS) estimators. The goal is to reduce variance by introducing bias. Papers published since then have built on these two ideas to work on different functional forms and improve performance. For example, the author of [2] has provided a review of Poisson regression. Moreover, recent developments in computing power introduced mathematical optimization to variable selection.

The aim of this paper is to review and propose methods to solve multicollinearity. The methods can be decided depending on the purpose of the regression, whether forecasting or analysis. Recent developments in machine learning and optimization have shown better results than conventional statistical methods. The pros and cons of the methods will be discussed in later sections. The paper is organized as follows. The multicollinearity phenomenon is explained in Section 2, including its effects and ways to measure it. Section 3 discusses the methods to reduce the effects of multicollinearity, including variable selection, modified estimators, and machine learning methods. Section 4 presents the concluding remarks.

2. What Is Multicollinearity?

Multicollinearity is a condition where there is an approximately linear relationship between two or more independent variables. This is a multiple linear regression model:

$$y = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p + e, \quad (1)$$

where y is the dependent variable, $x_1, \ldots, x_p$ represent the explanatory variable, β_0 is the constant term, $\beta_1, \ldots, \beta_p$ are the coefficients of the explanatory variable, and e is the error term. The error term is the difference between the observed and the estimated values. It is normally distributed with a mean of 0 and variance σ. In the presence of multicollinearity, x_1 may be linearly dependent on another explanatory variable such as x_2. The resulting model would be unreliable. The effects and problems are discussed in the following section.

For example, when using technical indicators in stock analysis. There will be a multicollinearity issue if the indicators measure the same type of information such as momentum [3]. The different indicators are all derived from the same series of closing prices in such a case. In the context of the stock market, data are handled differently from time-series data in other fields. This is due to the following few key reasons, according to The authors of [4]. The goal of compiling stock market data is to maximize profit and not reduce prediction error. Stock market data are highly time-variant, which means the output depends on the moment of the input. They are also dependent on indeterminate events. This means that the event that causes the response is not fixed.

2.1. Effects of Multicollinearity

According to [5], there are four main symptoms of multicollinearity. The first one is a large standard error of the coefficients. Next, the sign of a variable coefficient can be different from the theory. The explanation of the variable's effect on the output will be wrong or misleading. In addition, there will be a high correlation between the predictor variable and outcome, but the corresponding parameter is not statistically significant. The last symptom is that some correlation coefficients among predictor variables are large in relation to the explanatory power or R-Squared of the overall equation.

These are merely symptoms and do not guarantee the presence of multicollinearity. There are two major problems of multicollinearity. Estimates are unstable due to the interdependence of the variables and standard errors if the regression coefficient is large. This makes the estimates unreliable and therefore decreases their precision [6]. As two or more variables have linear relationships, the marginal impact of a variable is hard to measure. The model will have poor generalization ability and overfit the data. This means that it will perform poorly on data it has never seen.

2.2. Ways to Measure Multicollinearity

Previous literature found that there are four measurements of multicollinearity. The first detector of multicollinearity is a pairwise correlation using a correlation matrix. According to [7], a bivariate correlation of 0.8 or 0.9 is commonly used as a cut-off to indicate a high correlation between two regressors. However, the problem with this method is that the correlations do not necessarily mean multicollinearity as they are not the same. The most widely used indicator of multicollinearity is the Variation Inflation Factor (VIF) or Tolerance (TOL) [8]. The VIF is defined as

$$VIF_j = \frac{1}{\left(1 - R_j^2\right)}, \tag{2}$$

where R_j^2 is the coefficient of determination for the regression of x_j on the remaining variables. The VIF is the reciprocal of TOL. There is no formal value of VIF to determine the presence of multicollinearity, but a value of 10 and above often indicates multicollinearity [9].

Another method of measuring multicollinearity is using eigenvalues, which is from the Principal Component Approach (PCA). A smaller eigenvalue indicates a larger probability of the presence of multicollinearity. The fourth method is the Condition Index (CI). It is based on the eigenvalue. CI is the square root of the ratio between the maximum eigenvalue and each eigenvalue. According to [10], a CI of between 10 to 30 indicates moderate multicollinearity, while above 30 indicates severe multicollinearity.

VIF and CI 2 are commonly used treatments to determine the severity of the dataset before performing the methods to solve multicollinearity. It is important to note that the effectiveness of the two treatments in reducing multicollinearity is usually determined by comparing the root mean square error or out-sample forecast before and after treatments [11].

3. Reducing the Effects of Multicollinearity

Collecting more data is one of the simplest solutions to reduce the effects of multicollinearity because collinearity is more of a data problem than model specification problem. However, this is not always feasible, especially when research is undertaken using convenience sampling [1]. There is a cost associated with collecting more data. Furthermore, the quality of data collected might be compromised. Methods to eliminate multicollinearity by reducing the variances of regressor variances can be categorized into two methods: variable selection and modified estimates. Both methods can be applied at the same time. The detail of the variable selection and modified estimates methods are explained in the following sub-topics. Next, the machine learning approaches are also presented.

3.1. Variable Selection Methods

Researchers are mainly concerned about multicollinearity problems when forecasting with a linear regression model. Generally, researchers try to mitigate the effects of multicollinearity by using variable selection techniques so that a more reliable estimate of the parameter can be obtained [12]. These are commonly heuristic algorithms and rely on using indicators. The method can be completed by combining or eliminating variables. However, caution must be taken not to compromise the theoretical model to reduce multicollinearity. One of the earliest methods was stepwise regression. There are two basic processes, namely forward selection and backward elimination [13]. The forward selection method starts with an empty model and adds variables one at a time, while the backward elimination method starts with the full model with all available variables and drops them one by one. In each stage, they select the variable with the highest decrease in the residual sum of squares for forward selection or the lowest increase in the residual sum of squares for backward elimination.

However, there are some drawbacks to stepwise regression. According to the author of [14], it does not necessarily yield the best model in terms of the residual sum of squares

because of the order that these variables are added. This is especially true in the presence of multicollinearity. It is also not clear which of the two methods of stepwise regression is better. Furthermore, it also assumes there is a best equation when there can be equations with different variables that are equally as good. Another problem of the selection criterion is the computational effort required [15]. There are 2^k possible combinations for k independent variables. The amount of computation needed also increases exponentially with the total number of independent variables.

To reduce computation time, the authors of [16] therefore developed a more comprehensive method to fit the equation to the data. It uses a fractional factorial design with the statistical criteria, *Cp*, to avoid computing all the possible equations. It also works better on data with multicollinearity as the effectiveness of a variable is evaluated by its presence or absence in an equation. The *Cp* criterion was developed by the author of [17]. It provides a way to graphically compare different equations. The selection criterion *Cp* is as follows:

$$Cp = \frac{RSSp}{\hat{\sigma}^2} - (\, n - 2p \,), \tag{3}$$

where p is the number of variables, RSS, is the residual sum of squares for the regression being considered and $\hat{\sigma}^2$ is an estimate of σ^2; it is frequently the residual mean square from the complete regression. The model with a lower *Cp* is better.

Later, the authors of [18] proposed a more general selection criterion, *Sp* that has shown to outperform the *Cp* criterion. Methods that are based on the least square estimator such as the *Cp* criterion suffer in the presence of outliers and when the error variable deviates from normality. The *Sp* criterion solves this problem and can be used with any estimator of β without a need for modification. The *Sp* criterion is defined as follows:

$$S_p = \sum_{i=1}^{n} \left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2 / \sigma^2 - (k - 2p), \tag{4}$$

where k and p are the parameters of the full and subset model, respectively.

Information criteria provide an attractive way for model selection. Other criterions that are often used include, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), etc. [19]. According to [20], the difference between AIC and BIC is that BIC is consistent in selecting the model when the true model is under consideration. Meanwhile, AIC aims to minimize risk functions when the true model is not one of the candidates. The choice of criterion depends on the researcher and both AIC and BIC are suggested to be used together. Table 1 provides a summary of each stepwise feature selection and quality criterion.

Table 1. Stepwise feature selection and quality criterions.

Author	Year	Objective	Method	Pros	Cons
Ralston and Wilf [13]	1960	Develop a method for model selection	Forward selection and backward elimination	Simple to understand and use	Final model affected by order
Mallows [17]	1964	A criterion for subset selection	*Cp* criterion	Graphically compare quality between models	Suffers with outlier and non-normality
Gorman and Toman [16]	1966	Fractional factorial design to model selection	Fractional factorial design with the statistical criteria, *Cp*	Avoid computing all possible model	Heuristic technique
Kashid and Kulkarni [18]	2002	A more general selection criterion than *Cp* when least square is not best	*Sp* criterion	Applicable on any estimator	Computationally difficult and not consistent result
Misra and Yadav [21]	2020	Improve classification accuracy in small sample size	Recursive Feature Elimination with Cross-Validation	Does not delete the records	Evaluated on small sample size

The authors of [5] proposed the use of principal component analysis (PCA) as a solution to multicollinearity among predictor variables in a regression model. It is a statistical method to transform variables into new uncorrelated variables called principal components and reduce the number of predictive variables. Regression analysis is done using the principal components. The principal components are independent and thus satisfy the OLS assumption. The principal components are ranked according to the magnitude of the variance. This means that principal component 1 is responsible for a larger amount of variation than principal component 2. Therefore, PCA is useful for dimensionality reduction. Principal components with an eigenvalue near zero can be eliminated. This way the model is sparse while not dropping variables that might contain useful information.

The Partial Least Squares (PLS) method was developed by the author of [22]. It is a better alternative to multiple linear regression and PCA because it is more robust. The model parameters do not change by much when new samples are used. PLS is like PCA as it is also a dimension reduction technique. The difference is that it captures the characteristics of both X and Y instead of just X as does PCA. The PLS method works by iteratively extracting factors from X and Y while maximizing the covariance between the extracted factors. The PLS derives its usefulness from its ability to analyze noisy data with multicollinearity. This is because its underlying assumptions are much more realistic than traditional multiple linear regression. The authors of [23,24] compared the PLS method with the lasso and stepwise method and found it to be performing better.

A few journals have made comparisons among the techniques. The authors of [25] discussed and compared PCA and PLS as they are both dimension reduction methodologies. Both methods are used to convert highly correlated variables into independent variables and variable reduction. The methodology of PCA does not consider the relationship between the predictor variable and the response variable, while PLS does not. Therefore, the PCA is a dimension reduction technique that is unsupervised and PLS is a supervised technique. They also found that the predictive power of the principal components does not line up with the order. For example, the principal component 1 explains the change in response variable less than principal component 2. PLS is more efficient than PCA in this regard as it is a supervised technique. PLS is extracted based on significance and predictive power. The author of [26] compared partial least square regression (PLSR), ridge regression (RR), and principal component regression (PCR) using a simulation study. The study used MSE to compare the methods. They have found that when the number of independent variables increases, PLSR is the best. If the number of observations and the number of multicollinearities are large enough while the number of independent variables is small, RR has the smallest MSE.

Recent application of PLS is seen in the chaos phase modulation technique for underwater acoustic communication. The authors of [27] adopted a PLS regression into the chaos phase modulation communication to overcome the multicollinearity effect. They described PLS as a machine learning method that uses the training and testing processes simultaneously. The study found that this method effectively improves the communication signals. The authors compared it with two algorithms: the Time Reversal Demodulator and 3-layer Back Propagation Neural Network that does not perform feature analysis and relationship analysis. It shows that PLS regression has the best performance.

A multigene genetic programming was first developed by the authors of [28,29], who used this method to automate predictor selection to alleviate multicollinearity problems. The authors of [30] described a genetic algorithm-based machine learning approach to perform variable selection. The genetic algorithm is a general optimization algorithm based on concepts such as evolution and survival of the fittest. The model is initialized with creating a population with several individuals. Each individual is a different model. The genes of the model are features of the model. An objective function is used to determine the fitness of the models. In the next generation/iteration, the best model will be selected and have their genes "crossover". Some features of the parent model are combined. Mutation may also occur with some determined probability where the feature is reversed. According

to [30], this machine learning concept should be combined with a derivative based search algorithm for a hybrid model. This is because genetic algorithms are very good at finding generally good solutions but not good at finding local minima, such as derivative based search algorithms. Derivative based search algorithms can be performed after a certain amount of iteration of the genetic algorithm. Iterations are continued until no improvement in the fitness of the model is seen.

The authors of [31] proposed a quadratic programming approach to feature selection because previous methods do not consider the configuration of the dataset and therefore is not problem dependent. The aim of using quadratic programming is to maximize the number of relevant variables and reduce similar variables. The criterion Q that represents the quality of a subset of features a is presented in quadratic form. $Q(a) = a^T Qa - b^T a$, where $Q \in R^{nxn}$ is a matrix of pairwise predictor similarities, and $b \in R^n$ is a vector for relevance of the predictor to the target vector. The author suggested that the similarity between the features x_i and x_j and between x_i and y can be measured using Pearson's correlation coefficient [32] or the concept of mutual information [33]. However, these two methods do not directly capture feature relevance. The authors utilized a standard t-test to estimate the normalized significance of the features to account for it. This proposed method outperforms other feature selection methods, such as Stepwise, Ridge, Lasso, Elastic Net, LARS, and the genetic algorithm.

The authors of [34] presented the maximum relevance–minimum multicollinearity (MRmMC) method to perform variable selection and ranking. Its approach focusses on relevancy and redundancy as well. Relevancy refers to the relationship between features and the target variable, while redundancy is the multicollinearity between features. The main advantage of this paper over others is that it does not require any parameter tuning and is relatively easy to implement. Relevant features are measured with a correlation coefficient and redundancy with squared multiple correlation coefficient. A measure J that combines relevancy and multicollinearity is developed.

$$J(f_j) = \max_{\text{fj} \in \text{F} - \text{S}} [r^2_{qn}(f_j, c) - \sum_{i=1}^{k} sc(f_j, q_i)], \tag{5}$$

where r^2 is the correlation coefficient between feature f and target c. sc is the squared multiple correlation coefficient between feature f and its orthogonal transformed variable q. The first feature is selected using the optimization criteria V and the following are selected based on criterion J using a forward stepwise method. Although non-exhaustive, it is a very competent method for feature selection and reducing dimension.

The authors of [35] suggested that the mixed integer optimization (MIO) based approach to selecting variables has received increased attention with development in algorithms and hardware. They developed mixed integer quadratic optimization (MIQO) to eliminate multicollinearity in linear regression models. It adopts VIF as an indicator for detecting multicollinearity. Subset selection is performed subject to an upper bound constraint on VIF of each variable. It achieved higher R-Squared than heuristic-based methods such as stepwise selection. The solution is also computationally tractable and simpler to implement than cutting plane algorithm.

The authors of [36] proposed a profiled independence screening (PIS) method of screening for variables with high dimensionality and highly correlated predictors. It is built upon sure independence screening (SIS) [37]. Many variable selection methods developed before SIS do not work well in extremely high dimension data where predictors vastly outnumber the sample size. However, SIS may break down where the predictors are highly correlated, which resulted in the PIS. A factor profiling operator $Q(Z_I) = I_n - Z_I (Z_I^T Z_I)^{-1} Z_I^T$ is introduced to eliminate the correlation between predictors. The profiled data are applied to the SIS. $Z_I \in R^{nxd}$ is the latent factor matrix of X and d is the number of latent factors. Factor profiling is as follows:

$$Q(Z_I)y = Q(Z_I)X\beta + Q(Z_I)\varepsilon, \tag{6}$$

$Q(Z_I)y$ is the profiled response variable and the columns of $Q(Z_I)X$ are the profiled predictors. However, PIS may be misleading in a spiked population model. Preconditioned profiled independence screening (PPIS) solves this by using preconditioning and factor profiling. Two real data analyses show that PPIS has good performance.

Outlier detection is also a viable method for variable selection. Recently, projection pursuit was used to perform an outlier detection-based feature selection [37]. Projection pursuit aims to look for the most interesting linear projections. The author optimized it to find outliers. The method was found to be effective in improving classification tasks. However, it performs poorly when most features are highly correlated or when features are binary. Table 2 provides a summary of the findings for variable selection approaches.

Table 2. A summary of previous studies on variable selection.

Author	Year	Objective	Method	Pros	Cons
Wold [22]	1982	Creates new components using the relationship of predictor and response	Partial Least Square (PLS)	Supervised component extraction	Cannot exhibit significant non-linear characteristics
Lafi and Kanenee [5]	1992	Using PCA to perform regression	Principal component analysis (PCA)	Reduce dimensions	Does not account for relationship with response variable
Bies et al. [30]	2006	Genetic algorithm-based approach to model selection	Genetic algorithm	Less subjectivity on model	Not good in finding local minima
Katrutsa and Strijov [31]	2017	Quadratic programming approach	Quadratic programming	Investigates the relevance and redundancy of features	Cannot evaluate multicollinearity between quantitative and nominal random variable.
Senawi et al. [34]	2017	Feature selection and ranking	Maximum relevance-minimum multicollinearity (MRmMC)	Works well with classifying problems	Non-exhaustive
Tamura et al. [11]	2017	Mixed integer optimization	Mixed integer semidefinite optimization (MISDO)	Uses backward elimination to reduce computation	Only applies to low number of variables
Tamura et al. [35]	2019	Mixed integer optimization	Mixed integer quadratic optimization (MIQO)	Uses VIF as indicator	Only applies to low number of variables
Chen et al. [38]	2020	Combines the result of filter, wrapper, and embedded feature selection	Ensemble feature selection	Overcome local optima problem	Higher computation cost than single solution
Zhao et al. [36]	2020	Variable screening based on sure independence screening (SIS)	Preconditioned profiled independence screening (PPIS)	Variable screening in high dimensional setting	Require decorrelation of the predictors
Larabi-Marie-Sainte [39]	2021	Feature selection based on outlier detection	Projection Pursuit	Found outliers correlated with irrelevant features	Does not work well when features are noisy
Singh and Kumar [40]	2021	Creates new variables	Linear combination and ratio of independent variables	Does not remove any variables	Based on trial-and-error

The variable selection methods aim to reduce the number of variables to the few that are the most relevant. This may reduce the information gain from having more data

to work with. Furthermore, the modern optimization methods depend on subjectively determined indicators of relevance and similarity. This can be seen from [11] where the authors suggested other measures of multicollinearity for future research. It is therefore difficult to suggest which method is better without directly comparing performance on the same dataset. Better performance can also be due to the specific problem tested.

3.2. Modified Estimators Methods

Modified Estimators is another approach that use biased and shrunken estimators in exchange for lower variance and thus reduce overfitting [12]. The advantage is that the theoretical model is not compromised because of the dropping of variables. Its disadvantage is that the estimators are now biased. The most known method is the ridge regression developed by the author of [41]. This method adds a penalty term: the squared magnitude of the coefficient β to the loss function. The general equation of ridge regression is as follows:

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2, \tag{7}$$

The main issue with ridge regression is how to find ridge parameter λ. If λ is equal to zero, then the estimate will equal to the ordinary least square estimate. However, if the λ is too big, it will lead to an underfitting of the model. λ is selected by looking for the least increase in the root mean square error (RMSE) within an appropriate decrease in ridge variable inflation factors for each variable. A ridge trace is used to assist in this. It is a plot of coefficient $\hat{\beta}$ versus λ. It is used to pick the smallest λ at which the coefficients start to level off. Alternatively, a validation dataset is used, find λ that minimizes validation SSE. Identify λ such that the reduction in the variance term of the slope parameter is larger than the increment in its squared bias. The authors of [42] reviewed estimation methods for λ and new methods were suggested. A more recent paper proposed a Bayesian approach to solving the problem of finding the ridge parameter [43]. Simulation result showed that the approach is more robust and provide more flexibility in handling multicollinearity. Later, the authors of [44] proposed another way of solving the problem of finding the ridge parameter. Their generalized cross-validation approach to is able to find the global minimum.

More estimators have been developed from the ridge estimator. The authors of [45] used a jack-knife procedure to reduce the significant amount of bias of estimators from ridge regression. The author of [46] proposed a new class of estimator, the Liu estimator, based on the ridge estimator. It has the added advantage of a simple procedure to find the parameter λ. This is because the estimate is a linear function of λ. The author of [47] proposed the Liu-Type estimator. They found that the shrinkage of ridge regression is not effective when faced with severe multicollinearity. The Liu-Type estimator has a lower MSE when compared to ridge regression and fully addresses severe multicollinearity.

Since then, variations on ridge and Liu-type estimators have been created for use in different types of regression. The authors of [48] proposed a Liu-type estimator for binary logistic regression that is a generalization of the Liu-type estimator for a linear model. The authors of [49] stated that not much attention has been given to shrinkage estimators for generalized linear regression models, such as the Poisson regression model, logistic regression model, and negative binomial regression model. Therefore, they introduced a two-parameter shrinkage estimator for negative binomial models. It is a combination of the ridge estimator and Liu estimator. The authors of [50] modified the Jackknifed ridge regression estimator to form a Modified Jackknifed Poisson ridge regression estimator. The author of [2] reviewed the biased estimators in the Poisson regression model in the presence of multicollinearity. The regular maximum likelihood method in estimating regression coefficient is not reliable in the presence of multicollinearity. They compared the performance of four estimators in addition to the widely used ridge estimator and found that Liu-type estimators have superior performance over other methods in the Poisson regression model.

The authors of [51] proposed a partial ridge regression to solve three problem of regular ridge regression. Bias is applied to all variables regardless of the degree of multicollinearity in normal ridge regression. Stability is achieved at the cost of MSE and the selection method of λ is arbitrary. The proposed method applies the ridge parameter only to variables with a high degree of collinearity. This way the precision of the parameter estimator improves while retaining the MSE close to that of OLS. Estimates are closer to a true OLS estimate, β and overall variance is reduced significantly. It outperforms existing method in terms of bias, MSE, and relative efficiency.

The Lasso regression is a method developed by the author of [52] as a result of the problems of both stepwise regression and ridge regression. This problem is interpretability. Stepwise regression is interpretable, but the process is very discrete, as it is not known why variables are included or dropped from the model. Ridge regression is very good in multicollinearity due to the stability of the shrank coefficient. However, it does not reduce the coefficient to zero, therefore resulting in models that are hard to interpret. The Lasso is known as L1 regularization, while the ridge regression is known as L2 regularization. The main difference between the two is that Lasso reduces certain parameter estimates to zero. This serves to select variables as well. The equation is shown below:

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,, \tag{8}$$

The penalty term, absolute value of the coefficient β is added to the lost function. In this equation, Y is a (nx1) vector of response, X is a (nxp) matrix of predictor variables and β is a (px1) vector of unknown constants. As with ridge regression, as λ approaches zero, the equation becomes closer to the least square estimate. However, if the λ value is very large, the coefficient approaches zero. The Ridge regression shrinks the estimator but does nothing in variable selection, while Lasso achieves both. Due to this reason, Lasso is more desirable. It is more parsimonious and therefore better in explaining the relationship between independent and dependent variables.

When faced with multicollinearity, Ridge and Lasso perform differently. Ridge tends to spread the effect evenly and shrink the estimators of all the variables. While Lasso is unstable and tends to retain one of the variables and eliminates all the others. Lasso performs poorly in the case where the number of variables, p, is more than the number of observations, n. It selects at most n variables. When $n > p$, the performance of Lasso is not as good as Ridge regression. The authors of [53] proposed an elastic net that combines both Ridge and Lasso regression. It has the advantage of both the regularization methods, and it also shows grouping effect. The elastic net groups variables that are highly correlated together. It either drops or retains all of them together. Typically, cross-validation is used to choose the tuning parameter. It was originally used by [54]. In cross-validation, a subset of sample is a holdout in order to validate the performance.

The authors of [55] developed the algorithm least angle regression (LARs). It takes inspiration from Lasso and stagewise regression and aims to be a computationally simpler method. The LARs begins similarly to forward regression where it starts with all coefficients equal to zero and then adds the predictor most correlated with the response. The next variable has as much correlation as the current residuals. LARs proceeds equiangularly between the predictors, along the "least angle direction", until the next most correlated variable. The authors of [56] also improved upon the Lasso regression by using a mixed-integer programming approach. It eliminates structured noised and thus makes it perform better in a high dimensional environment where $p > n$. The authors of [57] further expanded on the idea and developed several penalized mixed-integer nonlinear programming models. The models are also solvable by a meta heuristic algorithm.

The authors of [58] introduced a strictly concave penalty function called modified log penalty. It is contrary to the strictly convex penalty of Elastic net. It is aimed at achieving a parsimonious model even under the effects of multicollinearity. Methods such as the Elastic

net tend to focus on the grouping effect which means that collinear variables are included together. Table 3 provides a summary of the findings for modified estimator approaches.

Table 3. A Summary of previous studies on modified estimators.

Author	Year	Objective	Method	Pros	Cons
Hoerl [41]	1962	Adds bias in exchange for lower variance	Ridge regression	Reduces overfitting	Introduces significant amount of bias
Singh et al. [45]	1986	Address significant amount of bias in ridge regression	Jack-knife procedure	Simple method to obtain confidence intervals for the regression parameters.	Larger variance than ridge regression
Liu [46]	1993	Simple procedure to find ridge parameter	Liu estimator	Ridge estimate is a linear function of ridge parameter	Does not work in severe multicollinearity
Tibshirani [52]	1996	Address interpretability of stepwise and ridge regression	Lasso regression	Reduces coefficient to zero	Worse performance than Ridge and does not work when $p > n$
Liu [47]	2003	Existing method does not work in severe multicollinearity	Liu-type estimator	Allows large shrinkage parameter	Two parameter estimation
Efron et al. [55]	2004	Computational simplicity	Least angle regression (LARs)	Computationally simpler Lasso	Very sensitive to the presence of outliers
Zou and Hastie [53]	2005	Combines Ridge and Lasso regression	Elastic net	Achieves grouping effect	No parsimony
Chandrasekhar et al. [51]	2016	Applies Ridge parameters only on variable with high collinearity	Partial ridge regression	More precise parameter estimates	Subjective measure of high collinearity
Assaf et al. [43]	2019	A conditionally conjugate prior for the biasing constant	Bayesian approach to finding ridge parameter	Produce a marginal posterior of parameter given the data	Only focus on getting a single parameter
Nguyen and Ng [58]	2019	Strictly concave penalty function	Modified log penalty	Parsimony variable selection under multicollinearity	No grouping effect
Kibria and Lukman [59]	2020	Alternative to the ordinary least squares estimator	Kibria–Lukman estimator	Outperforms Ridge and Liu-type regression	Results depends on certain conditions
Arashi et al. [60]	2021	High-dimensional alternative to Ridge and Liu	Two-parameter estimator	Has asymptotic properties	Lower efficiency in sparse model
Qaraad et al. [61]	2021	Tune parameter alpha of Elastic Net	Optimized Elastic Net	Effective with imbalanced and multiclass data	Accuracy metric not discussed

Modified estimators aim to improve the efficiency in parameter estimation in the presence of multicollinearity. This comes with a bias–variance trade-off. Researchers can select the methods based on their purpose such as grouping effect or parsimony. However, it can require extensive knowledge to know which one works better on the problem. For example, some methods are shown to work better in high or low dimensionality, degree of multicollinearity. Moreover, some methods are for linear regression and modifications need to be made for other functional form predictions or classification problem.

3.3. Machine Learning Methods

In this section, we attempt to present the overall state of the multicollinearity problem in machine learning and introduce interesting algorithms that deal with it implicitly. It is proven that a neural network is superior to traditional statistical models. The authors of [62] used a feed forward artificial neural network to model data with multicollinearity and found that it has much better performance in terms of RMSE compared to the traditional ordinary least square (OLS). This shows that machine learning methods with more complex architecture have the potential to produce much better estimates of the parameters than statistical methods. The authors of [51] provided reasons why a machine learning algorithm might be better. They have no requirement for assumptions about the function, can uncover complex patterns, and dynamically learn changing relationships.

Next, it is observed that variable selection methods have been applied in neural networks. The authors of [63] proposed a hybrid method that combines factor analysis and artificial neural network to combat multicollinearity. ANN is not able to perform variable selection, therefore PCA is used to extract components. ANN is then applied to the components. This method is named FA-ANN (factor analysis–artificial neural network). It is compared with regression analysis and genetic programming. FA-ANN has the best accuracy among them. The advantage of FA-ANN and genetic programming is that it is not based on any statistical assumptions, so it is more reliable and trustworthy. In addition, they can generalize over new sample data unlike regression analysis. However, they are considered a black-box model and are hard to interpret. A more recent version of this approach has been used in quality control research. The authors of [64] proposed a residual (r) control chart for data with multicollinearity. They suggested a neural network because a generalized linear model (GLM) may not work best in asymmetrically distributed data. They concluded that neural network model and functional PCA (FPCA) can deal with the high dimensional and correlated data.

Furthermore, regularization and penalty mechanisms can also be used to solve multicollinearity in machine learning models. For example, the Regularized OS-ELM algorithm [65], OS-ELM Time-varying (OS-ELM-TV) [66], Timeliness Online Sequential ELM algorithm [67], Least Squares Incremental ELM algorithm [68], and Regularized Recursive least-squares [69]. However, these mechanisms increase the computational complexity. For this reason, The authors of [70] proposed a method called the Kalman Learning Machine (KLM). It is an Extreme Learning Machine (ELM) that uses a Kalman filter to update the output weights of a Single Layer Feedforward Network (SLFN). A Kalman filter is an equation that can efficiently estimate the state of a process that minimizes mean squared error. The state is not updated in the learning stage as with the concept of ELM. The resulting model has shown to outperform basic machine learning models in prediction error (RMSE) and computing time. However, it requires manual optimization by humans. A constructive approach to building the model is suggested.

Although deep learning (DL) has emerged as an efficient method to automatically learn data representation without feature engineering, its discussion in terms of multicollinearity is very limited. Based on this motivation, our paper discussed the properties of neural networks, such as the convolutional neural network (CNN), recurrent neural network (RNN), attention mechanism, and graph neural network, before illustrating the example in mitigating the multicollinearity issue.

CNN is a neural network which was first introduced by the authors of [71] in the field of computer vision. It developed the concept of local receptive fields and shared weights to reduce the number of network parameter. It is very interesting in its way of addressing relationships between features. Traditional deep neural network suffers from booming parameter issues. CNN adopted multiple convolutional and pooling (subsampling) layers to detect the most representative features before connecting to a fully connected network for prediction. Specifically, the convolutional layer applied multiple feature extractors (filter) to detect the local features and produce its corresponding feature map to represent each local feature. The composition of multiple feature maps may represent the entire series. The

pooling layer is a dimensional reducing method to extract the most representative features and lower the noise. The generated features maps are likely to be independent of each other and potentially mitigate the multicollinearity problem. For example, The authors of [72] proposed the CNNpred framework to model the correlations among different variable to predict the stock market movement. Two variants of CNNpred, namely 2D-CNNpred and 3D-CNNpred, were introduced in their paper to extract combined features from a diverse set of input data. It comprises five major US stock market indices, currencies exchange rate, future contracts, commodities prices, treasury bill rates, etc. Their results showed a significant predictive improvement as compared to the state-of-the-art baseline. Another interesting study by the authors of [73] proposed to integrate the features learned from different representation of the same data to predict the stock market movement. They employed chart images (e.g., Candle bar, Line bar, and F-line bar) derived from stock prices as additional input to the prediction the SPDR S&P 500 ETF movement. The proposed model ensembled Long Short-Term Memory (LSTM) and CNN models to exploit their advantages in extracting temporal and image features, respectively. Thus, the result showed that the prediction error can be efficiently reduced by integrating the temporal and image features from the same data.

Other than feature maps, there is another influential development, namely the attention mechanism in the Recurrent Neural Network (RNN). RNN was first proposed by the author of [74] to process sequential information. Based on [75], the term "recurrent" explained the general architecture idea where a similar function applied on each element of the sequence and the computed output of the previous element will be aggregately retained over the internal memory of RNN until the end of sequence. Based on this, RNN enables compressing the information and producing a fixed-size vector to represent a sequence. The recurrence operation of RNN is advantageous in series data since the inherent information of a sequential can be effectively captured. Unlike CNN, RNN is more flexible to model a variable length of a sequence that can capture unbounded contextual information. However, the authors of [76] criticized that the recurrent-based model may be problematic in handling the long-range dependencies in data due to the memory compression issue in which the neural network struggles to compress all the necessary information from a long sequence input into a fixed-length vector. In order words, it is difficult to represent the entire input sequence without any information loss using the fixed-length vector. Despite the help of the gated activation function, the forgetting issues of RNN-based model becomes serious as the length of input sequence grows. Based on this, the attention mechanism was proposed to deal with the long-range dependencies issue by enabling the model to focus on the relevant part of input sequence when predicting a certain part of the output sequence.

According to [77], the attention mechanism was used to simulate visual attention where humans usually adjust their focal point over time to perceive a "high resolution" when focusing on a particular region of an image but perceive a "low resolution" for the surrounding image. Similarly, the attention mechanism enables the model to learn to assign different weights according to their contribution and may capture asymmetric influence between the features to mitigate the multicollinearity problem. For example, the authors of [78] proposed a CNN based on deep factorization machine and attention mechanism (FA-CNN) to enhance feature learning. In addition to capturing temporal influence, the attention mechanism enables modeling of the intraday interaction between the input features. The result showed a 7.38% improvement over LSTM in predicting stock movement.

Recently, there is another promising research to apply Graph Convolutional Networks (GCN) or graph embeddings in series data. Graph neural networks convert series data into a graph-structured data while enabling the model to capture the interconnectivity between the nodes. The interconnectivity or correlation modeling is relatively useful in reducing the multicollinearity effect. For example, the authors of [79] proposed the hierarchical graph attention network (HATS) to process the relational data for stock market prediction. Their study defined the stock market graph as a spatial–temporal graph where each individual

stock (company) is regarded as a node. Each node feature represented the current state of each company in response to its price movement and the state is dynamic over time. Based on this, HATS can selectively aggregate important information from various relation data to represent the company as a node. Thereafter, the model is trained to learn the interrelation between nodes before feeding into a task-specific layer for prediction. Table 4 provides a comprehensive summary of the machine learning approaches reviewed.

Table 4. A summary of machine learning approaches on solving multicollinearities.

Author	Year	Objective	Method
Huynh and Won [65]	2011	Multi-objective optimization function to minimize error	Regularized OS-ELM algorithm
Garg and Tai [63]	2012	Hybrid method of PCA and ANN	Factor analysis-artificial neural network (FA-ANN)
Ye et al. [66]	2013	Input weight that changes with time	OS-ELM Time-varying (OS-ELM-TV)
Gu et al. [67]	2014	Penalty factor in the weight adjustment matrix	Timeliness Online Sequential ELM algorithm
Guo et al. [68]	2014	Smoothing parameter to adjust output weight	Least Squares Incremental ELM algorithm
Hoseinzade and Haratizadeh [72]	2019	Model the correlation among different features from a diverse set of inputs	CNN-pred
Kim and Kim [73]	2019	Using features from different representation of same data to predicting the stock movement	LSTM-CNN
Nóbrega and Oliveira [70]	2019	Kalman filter to adjust output weight	Kalman Learning Machine (KLM)
Hua [80]	2020	Decision tree to select features	XGBoost
Obite et al. [62]	2020	Compare ANN and OLSR in presence of multicollinearity	Artificial neural network
Zhang et al. [78]	2021	Applied attention to capture the intraday interaction between input features	CNN-deep factorization machine and attention mechanism (FA-CNN)
Mahadi et al. [69]	2022	Regularization parameter that varies with time	Regularized Recursive least-squares

4. Conclusions

Most methods of solving for multicollinearity can be categorized as one of two. That is variable selection and modified estimators. This paper detailed the development of different methods over the years. Variable selection has the benefit of being simple to perform and can result in a sparse model. This makes it easy to interpret and does not overfit. The disadvantage is that the selection is very discretionary. There is also the underlying assumption that there is a best model when a model with different variables can be equally good. Recent papers on solving multicollinearity capitalizes on better computing power. The subset selection problem can be presented as an optimization problem where they search for the least redundant variable to reduce multicollinearity. In addition, they are also able optimize on the most relevant variables. They are based on criterions developed from previous literature that can represent relevance and redundancy. In this way, problems where there is high dimensionality (more variables than observations) can be handled.

Modifying the estimators is a very broad method and is more complex. There is different modified estimator for every functional form of the data. One of the main problems of modifying estimators is interpretability. It is very difficult to explain coefficients that are close to but not zero. However, it is more robust and performs better in the presence of multicollinearity. Some modified estimators even can perform variable selection.

The authors of [81] performed a stress test experiment on the following variable selection methods: Stepwise, Ridge, Lasso, Elastic Net, LARs, and Genetic algorithms. They compared their performance according to several quality measures on a few synthetic

datasets. The authors of [82] compared various statistical and machine learning methods. It is important to note that comparisons between methods have their drawback. For example, different tuning parameters can affect performances of the methods. The author of [14] considered domain knowledge in the studied field to be important in selecting variables as statistics alone is not enough in practice. All the methods are performed at a different degrees when dealing with different types of data.

Both variable selection and modified estimators can be used together. The number of features can be rapidly reduced to below the number of samples and then modified estimators can be applied. This can be seen in machine learning papers. The findings in this review paper are that variable selection drops variables and reduces information gain, while the multicollinearity measures to optimize are subjective. In addition, modified estimators have inconsistent performance depending on the data and are not able to be applied in every problem. The literature review also showed that machine learning algorithms are better than the simple OLS estimator in fitting data with multicollinearity. They do not need to have information on the relationships among the data or the distribution. This paper suggests that the relevancy and redundancy concept from feature selection can be adopted when training a machine learning model.

Author Contributions: J.Y.-L.C. and S.M.H.L. investigated the ideas, review the systems and methods, and wrote the manuscript. K.T.B. provided the survey studies and methods. W.K.C. conceived the presented ideas and wrote the manuscript with support from Y.-L.C., S.W.P.; and Z.-W.H. provided the suggestions on the experiment setup and provided the analytical results. J.Y.-L.C. and Y.-L.C. both provided suggestions on the research ideas, analytical results, wrote the manuscript, and provided funding supports. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Fundamental Research Grant Scheme provided by the Ministry of Higher Education of Malaysia under grant number FRGS/1/2019/STG06/UTAR/03/1. This work was also supported by the Ministry of Science and Technology in Taiwan, grants MOST-109-2628-E-027-004–MY3 and MOST-110-2218-E-027-004, and also supported by the Ministry of Education of Taiwan under Official Document No. 1100156712 entitled "The study of artificial intelligence and advanced semiconductor manufacturing for female STEM talent education and industry-university value-added cooperation promotion".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: No potential conflict of interest was reported by the authors.

References

1. Schroeder, M.A.; Lander, J.; Levine-Silverman, S. Diagnosing and dealing with multicollinearity. *West. J. Nurs. Res.* **1990**, *12*, 175–187. [CrossRef] [PubMed]
2. Algamal, Z.Y. Biased estimators in Poisson regression model in the presence of multicollinearity: A subject review. *Al-Qadisiyah J. Adm. Econ. Sci.* **2018**, *20*, 37–43.
3. Bollinger, J. Using bollinger bands. *Stock. Commod.* **1992**, *10*, 47–51.
4. Iba, H.; Sasaki, T. Using genetic programming to predict financial data. In Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406), Washington, DC, USA, 6–9 July 1999; pp. 244–251.
5. Lafi, S.; Kaneene, J. An explanation of the use of principal-components analysis to detect and correct for multicollinearity. *Prev. Vet. Med.* **1992**, *13*, 261–275. [CrossRef]
6. Alin, A. Multicollinearity. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 370–374. [CrossRef]
7. Mason, C.H.; Perreault, W.D., Jr. Collinearity, power, and interpretation of multiple regression analysis. *J. Mark. Res.* **1991**, *28*, 268–280. [CrossRef]
8. Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W. *Applied Linear Statistical Models*; WCB McGraw-Hill: New York, NY, USA, 1996.
9. Weisberg, S. *Applied Linear Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2005; Volume 528.
10. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
11. Tamura, R.; Kobayashi, K.; Takano, Y.; Miyashiro, R.; Nakata, K.; Matsui, T. Best subset selection for eliminating multicollinearity. *J. Oper. Res. Soc. Jpn.* **2017**, *60*, 321–336. [CrossRef]

12. Askin, R.G. Multicollinearity in regression: Review and examples. *J. Forecast.* **1982**, *1*, 281–292. [CrossRef]
13. Ralston, A.; Wilf, H.S. *Mathematical Methods for Digital Computers*; Wiley: New York, NY, USA, 1960.
14. Hamaker, H. On multiple regression analysis. *Stat. Neerl.* **1962**, *16*, 31–56. [CrossRef]
15. Hocking, R.R.; Leslie, R. Selection of the best subset in regression analysis. *Technometrics* **1967**, *9*, 531–540. [CrossRef]
16. Gorman, J.W.; Toman, R. Selection of variables for fitting equations to data. *Technometrics* **1966**, *8*, 27–51. [CrossRef]
17. Mallows, C. *Choosing Variables in a Linear Regression: A Graphical Aid*; Central Regional Meeting of the Institute of Mathematical Statistics: Manhattan, KS, USA, 1964.
18. Kashid, D.; Kulkarni, S. A more general criterion for subset selection in multiple linear regression. *Commun. Stat.-Theory Methods* **2002**, *31*, 795–811. [CrossRef]
19. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
20. Vrieze, S.I. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol. Methods* **2012**, *17*, 228. [CrossRef] [PubMed]
21. Misra, P.; Yadav, A.S. Improving the classification accuracy using recursive feature elimination with cross-validation. *Int. J. Emerg. Technol.* **2020**, *11*, 659–665.
22. Wold, H. Soft modeling: The basic design and some extensions. *Syst. Under Indirect Obs.* **1982**, *2*, 343.
23. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
24. Chong, I.-G.; Jun, C.-H. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103–112. [CrossRef]
25. Maitra, S.; Yan, J. Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Appl. Multivar. Stat. Models* **2008**, *79*, 79–90.
26. Onur, T. A Comparative Study on Regression Methods in the presence of Multicollinearity. *İstatistikçiler Derg. İstatistik Ve Aktüerya* **2016**, *9*, 47–53.
27. Li, C.; Wang, H.; Wang, J.; Tai, Y.; Yang, F. Multicollinearity problem of CPM communication signals and its suppression method with PLS algorithm. In Proceedings of the Thirteenth ACM International Conference on Underwater Networks & Systems, Shenzhen, China, 3–5 December 2018; pp. 1–5.
28. Willis, M.; Hiden, H.; Hinchliffe, M.; McKay, B.; Barton, G.W. Systems modelling using genetic programming. *Comput. Chem. Eng.* **1997**, *21*, S1161–S1166. [CrossRef]
29. Castillo, F.A.; Villa, C.M. Symbolic regression in multicollinearity problems. In Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation, Washington, DC, USA, 25–29 June 2005; pp. 2207–2208.
30. Bies, R.R.; Muldoon, M.F.; Pollock, B.G.; Manuck, S.; Smith, G.; Sale, M.E. A genetic algorithm-based, hybrid machine learning approach to model selection. *J. Pharmacokinet. Pharmacodyn.* **2006**, *33*, 195. [CrossRef] [PubMed]
31. Katrutsa, A.; Strijov, V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Syst. Appl.* **2017**, *76*, 1–11. [CrossRef]
32. Hall, M.A. *Correlation-Based Feature Selection for Machine Learning*; The University of Waikato: Hamilton, New Zealand, 1999.
33. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
34. Senawi, A.; Wei, H.-L.; Billings, S.A. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. *Pattern Recognit.* **2017**, *67*, 47–61. [CrossRef]
35. Tamura, R.; Kobayashi, K.; Takano, Y.; Miyashiro, R.; Nakata, K.; Matsui, T. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. *J. Glob. Optim.* **2019**, *73*, 431–446. [CrossRef]
36. Zhao, N.; Xu, Q.; Tang, M.L.; Jiang, B.; Chen, Z.; Wang, H. High-dimensional variable screening under multicollinearity. *Stat* **2020**, *9*, e272. [CrossRef]
37. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **2008**, *70*, 849–911. [CrossRef]
38. Chen, C.W.; Tsai, Y.H.; Chang, F.R.; Lin, W.C. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Syst.* **2020**, *37*, e12553. [CrossRef]
39. Larabi-Marie-Sainte, S. Outlier Detection Based Feature Selection Exploiting Bio-Inspired Optimization Algorithms. *Appl. Sci.* **2021**, *11*, 6769. [CrossRef]
40. Singh, S.G.; Kumar, S.V. Dealing with Multicollinearity problem in analysis of side friction characteristics under urban heterogeneous traffic conditions. *Arab. J. Sci. Eng.* **2021**, *46*, 10739–10755. [CrossRef]
41. Horel, A. Applications of ridge analysis toregression problems. *Chem. Eng. Progress.* **1962**, *58*, 54–59.
42. Duzan, H.; Shariff, N.S.B.M. Ridge regression for solving the multicollinearity problem: Review of methods and models. *J. Appl. Sci.* **2015**, *15*, 392–404. [CrossRef]
43. Assaf, A.G.; Tsionas, M.; Tasiopoulos, A. Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression. *Tour. Manag.* **2019**, *71*, 1–8. [CrossRef]
44. Roozbeh, M.; Arashi, M.; Hamzah, N.A. Generalized cross-validation for simultaneous optimization of tuning parameters in ridge regression. *Iran. J. Sci. Technol. Trans. A Sci.* **2020**, *44*, 473–485. [CrossRef]
45. Singh, B.; Chaubey, Y.; Dwivedi, T. An almost unbiased ridge estimator. *Sankhyā Indian J. Stat. Ser. B* **1986**, *48*, 342–346.

46. Kejian, L. A new class of biased estimate in linear regression. *Commun. Stat.-Theory Methods* **1993**, *22*, 393–402. [CrossRef]
47. Liu, K. Using Liu-type estimator to combat collinearity. *Commun. Stat.-Theory Methods* **2003**, *32*, 1009–1020. [CrossRef]
48. Inan, D.; Erdogan, B.E. Liu-type logistic estimator. *Commun. Stat.-Simul. Comput.* **2013**, *42*, 1578–1586. [CrossRef]
49. Huang, J.; Yang, H. A two-parameter estimator in the negative binomial regression model. *J. Stat. Comput. Simul.* **2014**, *84*, 124–134. [CrossRef]
50. Türkan, S.; Özel, G. A new modified Jackknifed estimator for the Poisson regression model. *J. Appl. Stat.* **2016**, *43*, 1892–1905. [CrossRef]
51. Chandrasekhar, C.; Bagyalakshmi, H.; Srinivasan, M.; Gallo, M. Partial ridge regression under multicollinearity. *J. Appl. Stat.* **2016**, *43*, 2462–2473. [CrossRef]
52. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
53. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
54. Mosier, C.I. Problems and designs of cross-validation 1. *Educ. Psychol. Meas.* **1951**, *11*, 5–11. [CrossRef]
55. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [CrossRef]
56. Roozbeh, M.; Babaie–Kafaki, S.; Aminifard, Z. A nonlinear mixed–integer programming approach for variable selection in linear regression model. *Commun. Stat.-Simul. Comput.* **2021**, 1–12. [CrossRef]
57. Roozbeh, M.; Babaie-Kafaki, S.; Aminifard, Z. Improved high-dimensional regression models with matrix approximations applied to the comparative case studies with support vector machines. *Optim. Methods Softw.* **2022**, 1–18. [CrossRef]
58. Nguyen, V.C.; Ng, C.T. Variable selection under multicollinearity using modified log penalty. *J. Appl. Stat.* **2020**, *47*, 201–230. [CrossRef]
59. Kibria, B.; Lukman, A.F. A new ridge-type estimator for the linear regression model: Simulations and applications. *Scientifica* **2020**, *2020*, 9758378. [CrossRef]
60. Arashi, M.; Norouzirad, M.; Roozbeh, M.; Mamode Khan, N. A High-Dimensional Counterpart for the Ridge Estimator in Multicollinear Situations. *Mathematics* **2021**, *9*, 3057. [CrossRef]
61. Qaraad, M.; Amjad, S.; Manhrawy, I.I.; Fathi, H.; Hassan, B.A.; El Kafrawy, P. A hybrid feature selection optimization model for high dimension data classification. *IEEE Access* **2021**, *9*, 42884–42895. [CrossRef]
62. Obite, C.; Olewuezi, N.; Ugwuanyim, G.; Bartholomew, D. Multicollinearity effect in regression analysis: A feed forward artificial neural network approach. *Asian J. Probab. Stat.* **2020**, *6*, 22–33. [CrossRef]
63. Garg, A.; Tai, K. Comparison of regression analysis, artificial neural network and genetic programming in handling the multicollinearity problem. Proceedings of International Conference on Modelling, Identification and Control, Wuhan, China, 24–26 June 2012; pp. 353–358.
64. Kim, J.-M.; Wang, N.; Liu, Y.; Park, K. Residual control chart for binary response with multicollinearity covariates by neural network model. *Symmetry* **2020**, *12*, 381. [CrossRef]
65. Huynh, H.T.; Won, Y. Regularized online sequential learning algorithm for single-hidden layer feedforward neural networks. *Pattern Recognit. Lett.* **2011**, *32*, 1930–1935. [CrossRef]
66. Ye, Y.; Squartini, S.; Piazza, F. Online sequential extreme learning machine in nonstationary environments. *Neurocomputing* **2013**, *116*, 94–101. [CrossRef]
67. Gu, Y.; Liu, J.; Chen, Y.; Jiang, X.; Yu, H. TOSELM: Timeliness online sequential extreme learning machine. *Neurocomputing* **2014**, *128*, 119–127. [CrossRef]
68. Guo, L.; Hao, J.-h.; Liu, M. An incremental extreme learning machine for online sequential learning problems. *Neurocomputing* **2014**, *128*, 50–58. [CrossRef]
69. Mahadi, M.; Ballal, T.; Moinuddin, M.; Al-Saggaf, U.M. A Recursive Least-Squares with a Time-Varying Regularization Parameter. *Appl. Sci.* **2022**, *12*, 2077. [CrossRef]
70. Nobrega, J.P.; Oliveira, A.L. A sequential learning method with Kalman filter and extreme learning machine for regression and time series forecasting. *Neurocomputing* **2019**, *337*, 235–250. [CrossRef]
71. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
72. Hoseinzade, E.; Haratizadeh, S. CNNpred: CNN-based stock market prediction using a diverse set of variables. *Expert Syst. Appl.* **2019**, *129*, 273–285. [CrossRef]
73. Kim, T.; Kim, H.Y. Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. *PLoS ONE* **2019**, *14*, e0212320. [CrossRef] [PubMed]
74. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
75. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
76. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
77. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
78. Zhang, X.; Liu, S.; Zheng, X. Stock Price Movement Prediction Based on a Deep Factorization Machine and the Attention Mechanism. *Mathematics* **2021**, *9*, 800. [CrossRef]

79. Kim, R.; So, C.H.; Jeong, M.; Lee, S.; Kim, J.; Kang, J. Hats: A hierarchical graph attention network for stock movement prediction. *arXiv* **2019**, arXiv:1908.07999.
80. Hua, Y. An efficient traffic classification scheme using embedded feature selection and lightgbm. In Proceedings of the Information Communication Technologies Conference (ICTC), Nanjing, China, 29–31 May 2020; pp. 125–130.
81. Katrutsa, A.; Strijov, V. Stress test procedure for feature selection algorithms. *Chemom. Intell. Lab. Syst.* **2015**, *142*, 172–183. [CrossRef]
82. Garg, A.; Tai, K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *Int. J. Model. Identif. Control* **2013**, *18*, 295–312. [CrossRef]