Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 2308541, 8 pages
https://doi.org/10.1155/2023/2308541

Research Article

Inverse Matrix Problem in Regression for High-Dimensional Data Sets

Namra Shakeel and Tahir Mehmood

School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Correspondence should be addressed to Tahir Mehmood; tahime@gmail.com

Received 19 December 2022; Revised 30 January 2023; Accepted 3 February 2023; Published 23 February 2023

Academic Editor: Taoreed Owolabi

Forhigh-dimensional chemometric data, the inverse matrix $(X^tX)^{-1}$ problem in regression models is a difficulty. Multicollinearity and identification result from the inverse matrix problem. The usage of the least absolute shrinkage and selection operator (LASSO) and partial least squares are two existing ways of dealing with the inverse matrix problem (PLS). For regressing the chemometric data sets, we used extended inverse and beta cube regression. The existing and proposed methods are compared over near-infrared spectra of biscuit dough and Raman spectra analysis of contents of polyunsaturated fatty acids (PUFA). For reliable estimation, Monte Carlo cross-validation has been used. The proposed methods outperform based on the root mean square error, indicating that cube regression and inverse regression are reliable and can be used for diverse high-dimensional data sets.

1. Introduction

Predictions have long been a key component of modern data science, whether in statistical analysis or machine learning. Modern technology allows for massive data expansion, yet this data frequently contains meaningless information, making prediction difficult. Researchers are employing novel methodologies and algorithms to extract information and build robust prediction models. Predictor variables that are either directly or indirectly related to other predictor factors are commonly included in such models. Multiple linear regression is widely employed in modern research fields such as chemometrics, econometrics, and bioinformatics [1]. Multiple linear regression models determine the relationship between multiple independent variables $X_1, X_2, X_3, \ldots, X_p$ i.e., X_{nxp} and dependent variable Y. It can be written as $Y = X\beta + \mu$, where β is the vector of model coefficients estimated by ordinary least square $\widehat{\beta} = (X^tX)^{-1}X^tY$ that represents the relation between response and predictors while μ is the error term in the model.

Certain traits or assumptions are associated with multivariate regression models for prediction. The correlation between predictor factors and the increased number of predictor variables are two of the characteristics. In presence of multicollinearity and a larger number of predictors, the inverse matrix that is $(X^tX)^{-1}$ does not hold. In chemometrics, most of the data sets faced such issues. For instance, near-infrared (NIR) and Raman spectroscopy calibrations of chemical components in food samples [2], the fatty acid composition and quantities of major constituents in a complex food model system [3] by using Raman spectroscopy. NIR is a vibrational spectroscopy method based on overtones and combinations of basic vibrational modes, whereas Raman is a vibrational spectroscopy technique based on fundamental stretching and deformation modes [4]. In comparison to Raman, the latter technique's spectral bands are usually wider, giving NIR a poor chemical selectivity. The Raman and near-infrared spectroscopies are both appropriate for food analysis that is quick and useful. Here, measurements are feasible with fiber optics [5]. Both techniques conduct the qualitative, quantitative, and structural information about the samples [6]. For solving the inverse matrix problem penalized regression models are being used. Examples of the penalized regression models

include ridged regression [7] and least absolute selection and shrinkage operator (LASSO) [8]. An alternative way is to use partial least squares [9] which utilizes the latent variables. Another dimension is to use the Moore–Penrose generalized inverse [10] for solving the inverse matrix issue. We have introduced the cube penalized regression and have reused the generalized inverse in multiple linear regression. Moore–Penrose generalized inverse is a method for solving singular matrices of high-dimensional regression data. Like other well-known LASSO and ridge regression methods, the concept of beta cube regression is established by considering the cube of the penalty parameter. These proposed methods are compared with reference methods i.e., LASSO and PLS over chemometric high-dimensional data [11–13]. Even though there is a vast literature on the topic of collinearity in linear regression, there is still a need to enhance the outcomes of traditional regression methods by incorporating new regression approaches. Newly discovered methods can save time by efficiently executing programs. Furthermore, these techniques can produce precise and effective outcomes.

2. Materials

Two data sets have been chosen for comparison. These data sets are then subdivided into each response, and all regression techniques will be run independently to each response variable with independent variables.

2.1. NIR (Near-Infrared) Spectra of Biscuit Dough. The data set contains 700 NIR spectra wavelengths (1100–2498 nm in 2 nm increments) that have been utilized as predictor variables. Fat, sugar, flour, and water yield percentages are arranged into four response variables. As a univariate response, the assessment of each answer is predicted independently [1, 14].

Figure 1 represents the flow for getting data from NIR spectra [15].

2.2. Raman Spectra Analysis of Contents of PUFA. The fatty acid composition across the 69 samples in the data set, when using Raman spectroscopy to extract specific chemical information from small components in meals [16]. The fatty acid content in this data set is provided as a percentage of the total sample weight and total fat content. As predictors, the samples produced 1096 wavelength variables. Figure 2 depicts the procedure for obtaining data from Raman spectra.

3. Methods

We have considered two reference methods LASSO and PLS while we have proposed two methods beta cube regression and generalized inverse regression.

3.1. Least Absolute Shrinkage and Selection Operator. The LASSO operator stands for least absolute shrinkage and selection operator. It has been introduced by Santosa and Symes in 1986 [17]. It came into prominence in 2006 by Tibshirani [18]. This is the type of regression model that executes variable selection as well as regularization to increase the estimation of accuracy. LASSO shrink β coefficients exactly equal to zero. LASSO is also called the L_1 norm [19]. The LASSO estimator of $\widehat{\beta}$ can be written as

$$\widehat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y - X\beta)^2 \text{ subject to } \sum_{j=1}^{p} |\beta|_1 \leq t, \tag{1}$$

where t is the penalty on L_1 norm. Equation 1 can also be written in the following form:

$$\widehat{\beta}_{LASSO} = \operatorname{argmin}_{\beta} \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta|_1. \tag{2}$$

Here, L_1 norm and L_2 norm are defined by $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$ and $\|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2$. There is one-to-one correspondence in t and λ. This relation is because of duality. Thus, for every $t \geq 0$, there exists $\lambda \geq 0$ such that both problems play the same role [20]. The selection of λ can be done by cross validation. If $\lambda = 0$, LASSO estimator behaves similarly to the ordinary least square. If λ value rises, a number of nonzero $\widehat{\beta}$ coefficients decreases and if λ value approaches to ∞ , then $\widehat{\beta}$ becomes zero and LASSO provides null model [21]. Due to the nondifferentiable objective function, the LASSO does not provide a closed-form solution to the problem. Still, there is a possibility of obtaining closed-form by adding a soft threshold operator. That is how the soft-thresholding operator is defined for LASSO regression.

$$\begin{aligned} \operatorname{sign}_{\lambda}(x) &= x + \lambda, \text{ if } x < -\lambda, \\ &0, \text{ if } |x| \leq \lambda, \\ &x - \lambda, \text{ if } x > \lambda. \end{aligned} \tag{3}$$

The $\widehat{\beta}$ can be written as

$$\begin{aligned} \widehat{\beta} &= \frac{1}{n}\left(X^t Y\right) - \frac{\lambda}{2}\operatorname{sign}(\beta), \\ \widehat{\beta}_{\text{LASSO}} &= \frac{1}{n}\left(X^t Y\right)_j + \frac{\lambda}{2}, \text{if } \frac{1}{n}\left(X^t Y\right)_j < -\frac{\lambda}{2}, \\ &0, \text{if } \frac{1}{n}\left|\left(X^t Y\right)_j\right| \leq \frac{\lambda}{2}, \\ &\frac{1}{n}\left(X^t Y\right)_j - \frac{\lambda}{2}, \text{if } \frac{1}{n}\left(X^t Y\right)_j > \frac{\lambda}{2}. \end{aligned} \tag{4}$$

3.2. Partial Least Square (PLS) Regression. PLS is a substitute of a multiple linear regression model [22]. It is an iterative procedure. In PLS, the objective is to optimize the covariance between X and Y. The PLS regression can be written as $\widehat{\beta}_{PLS} = W(P^t W)^{-1} Q$. Here, P is the X-loading, Q is the Y-loading, and W is the loading weights. PLS even performs when data is noisy and missing.

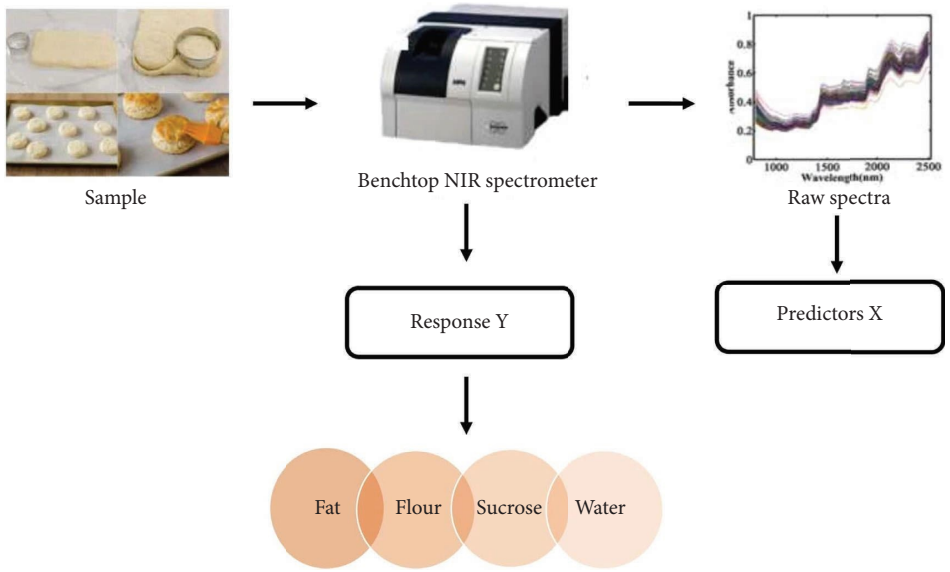

Figure 1: This figure represents the NIR spectra of biscuit dough.

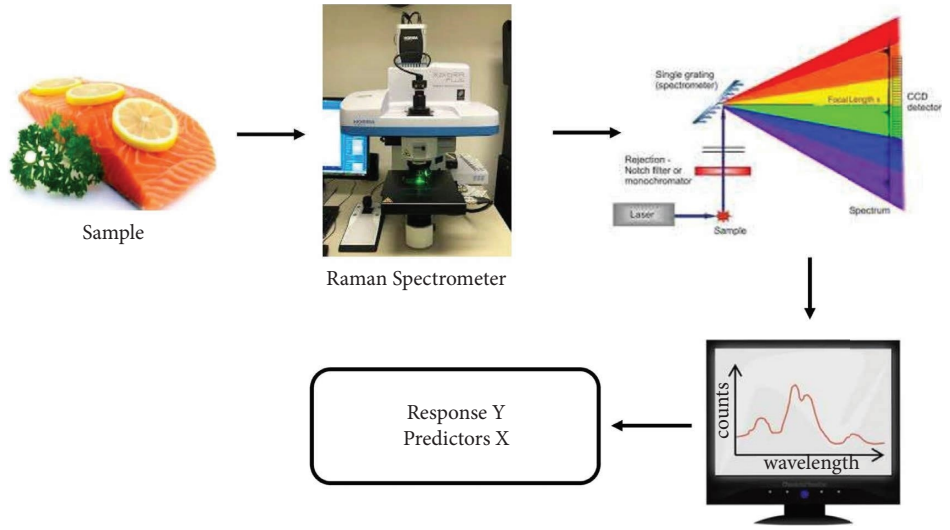

Figure 2: This figure represents the Raman spectroscopy of fatty acids.

3.3. Generalized Inverse Regression. For A is a $n \times p$ matrix of rank $\leq p$, then A^{-} of dimension $p \times n$ is called generalized inverse matrix $AA^{-}A = A$.

The generalized inverse is not unique in general, but it always exists. As the inverse of a matrix does not exist if its determinant becomes zero. The solution to finding the inverse of such a matrix is presented by an American mathematician, Moore in 1935, and later in 1955, a scientist named Penrose developed a Moore inverse in a different method [10]. Hence, is called Moore–Penrose Inverse. Although Moore–Penrose generalized inverse is not unique in the case of the nonsquare and singular matrix, it provides accurate results in minimum time for regression data sets. In the meanwhile, an author named Rao contributed a computational method of a singular matrix called pseudo inverse and used it to solve the least square theory to know estimators of linear equations [23]. Finding the system of linear equations is one of the most common uses of the generalized inverse. Let $Y = X\beta$ then $\beta = X^{-1}Y$, it is only true if X is invertible matrix. If X is nonsquare, singular matrix, then its Moore–Penrose generalized inverse can be represented as

$$\widehat{\beta}_{\text{ginv}} = \text{ginv}\left(X^{t}X\right)X^{t}Y, \tag{5}$$

where "ginv" presents the generalized inverse.

3.4. Beta Cube Regression. The concept of beta cube regression is basically taken from [24] by considering the $L3$ penalty in ordinary least squares. By considering some conditions, this regression method can be applied to real-life data sets. The regression coefficients are estimated using this approach by solving the following constraint, expressed as

$$\widehat{\beta}_{\text{Cube}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y - X\beta)^2 \text{ subject to } \sum_{j=1}^{p} |\beta|^3 \leq t. \quad (6)$$

Equation 6 can also be described as

$$\widehat{\beta}_{\text{Cube}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y - X\beta)^2 + \lambda \sum_{j=1}^{p} |\beta|^3. \quad (7)$$

Equating the derivative of equation 7, we will obtain the following equation:

$$\widehat{\beta}_{\text{Cube}} = \left(2X^t X + 3\lambda\beta\right)^{-1} 2X^t Y. \quad (8)$$

β vector comes within the $\widehat{\beta}$ after the derivation of constraint. β vector in the previous equations can first be generated randomly by normal random distribution and against each value of λ; $\widehat{\beta}$ is computed. Then, from all computed $\widehat{\beta}$, one optimal $\widehat{\beta}$ is selected. This optimal $\widehat{\beta}$ is generated again and again by applying a loop in it to get the best and minimum $\widehat{\beta}$.

4. Model Building and Comparison

For model building, the model parameters are required to tune. For this, we have used cross validation [25, 26]. Moreover, cross validation is used for the reliable comparison of reference regression models and proposed regression models.

4.1. Cross Validation. Cross validation (CV) is a fundamental approach for verifying the dependability of a regression model. The fundamental idea underlying CV is to assess a model's prediction performance using data that was not used to develop the model. Typically, technique performance is evaluated using new data. To eliminate this reliance, the data set is divided into training and test sets numerous times. The model parameters for all λ possibilities using the training data are derived for each split, and the estimated parameters are evaluated on the corresponding test set. The penalty parameter that performs best (in specific respects) across all train sets is then selected [28].

4.2. Monte Carlo Cross Validation. Data splitting can be carried out by Monte Carlo cross validation. In Monte Carlo cross validation, each data point is tested arbitrary times and partitions can be possible many times, and this method's result is high bias but low in variance. It splits the training data at random (maybe 70–30 percent, or 60–40 percent). Data are split randomly to avoid overestimating or underestimating the results [29].

4.3. Selection of Penalty Parameter. The penalized term in regression is determined by a tuning parameter λ, also known as a penalty parameter. When data values are shrunk towards a central point, such as the mean, it refers to the degree of shrinking that happens.

4.4. Performance Estimation. We employed root mean square error (RMSE), a method for measuring prediction quality, to estimate the performance of regression models. Using Euclidean distance, it predicts how far estimations differ from actual values.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n}}, \quad (9)$$

where n is sample size, Y_i is the i^{th} actual response and $\widehat{Y}_i$ is the i^{th} predicted response [30].

5. Results and Discussion

We have considered 2 data sets NIR spectra of biscuit dough and Raman spectra analysis of contents of PUFFA for modeling and comparison purposes. The descriptive summary of these data sets is presented in Tables 1 and 2. Notably, the NIR data set has 4 response variables fat, flour, sucrose, and water while the Raman data set has 2 response variables weight and fat.

In Figure 3, the suggested technique beta cube and generalized inverse approaches are compared to reference methods PLS and LASSO for prediction capabilities. The fact that the RMSE of the generalized inverse is almost equal to zero for cookie dough components demonstrates that it works well for each data set. In comparison to PLS, the beta cube performs best, with an RMSE value between 0 and 0.25. LASSO has the lowest RMSE among all approaches for fat and flour data sets, but it does not provide adequate results for sugar and water data sets. For all data sets of biscuit dough, in Figure 4 the generalized inverse is performing well for threshold = 0. This means for threshold = 0, the models are including all the $\widehat{\beta}$ coefficients. Beta Cube is picking different penalty parameters between 0.1 and 1 and thus getting the mean value of λ around 0.5. This means the model is picking approximately half of the relevant $\widehat{\beta}$ coefficients. The PLS model is considering optimal components in between the range 0.1 to 0.8 and providing a mean value of thresholds around 0.45. LASSO is predicting suitable results for threshold values equal to 0.1 and 0.2 for fat and flour data sets. For sucrose and water data sets, LASSO is not performing well even at a threshold value of 1.

Table 1: A descriptive description of the NIR biscuit dough data is provided.

Responses	Fat	Flour	Sucrose	Water
Mean	18.31	16.59	48.98	14.19
Variance	3.87	15.24	7.46	2.20
Training data	50	50	50	50
Testing data	22	22	22	22

Table 2: The content of polyunsaturated fatty acids is described in detail.

Responses	Sample weight	Fat content
Mean	4.39	33.64
Variance	7.89	251.85
Training data	48	48
Testing data	21	21

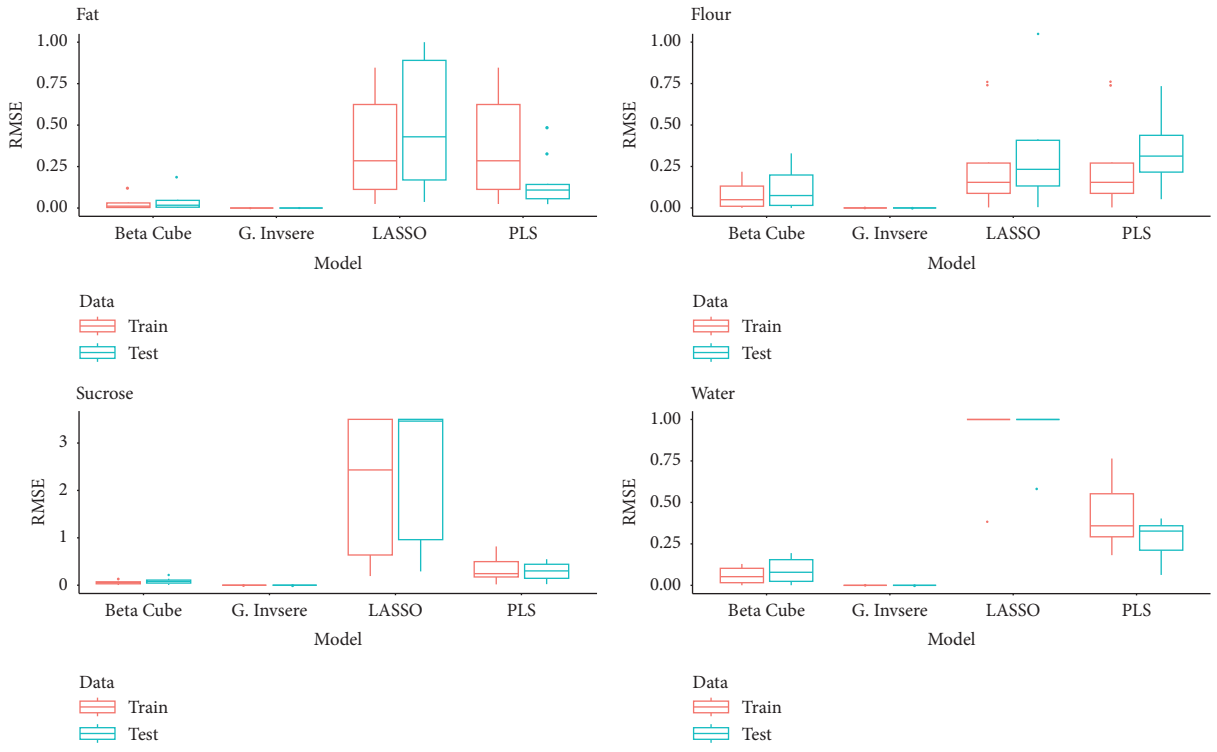

Figure 3: The RMSE of components of biscuit dough for each prediction method is presented.

Other methodologies' prediction accuracy is determined using PLS and LASSO algorithms. The RMSE of the generalized inverse for fatty acid is almost equal to zero, suggesting that it performs best for each data set, as shown in 5. In comparison to PLS, the beta cube performs best, with an RMSE value between 0 and 0.15. When compared to all other techniques, LASSO delivers the lowest RMSE for these data sets.

For all data sets of fatty acid, in Figures5, the generalized inverse is performing well for threshold = 0. This means for

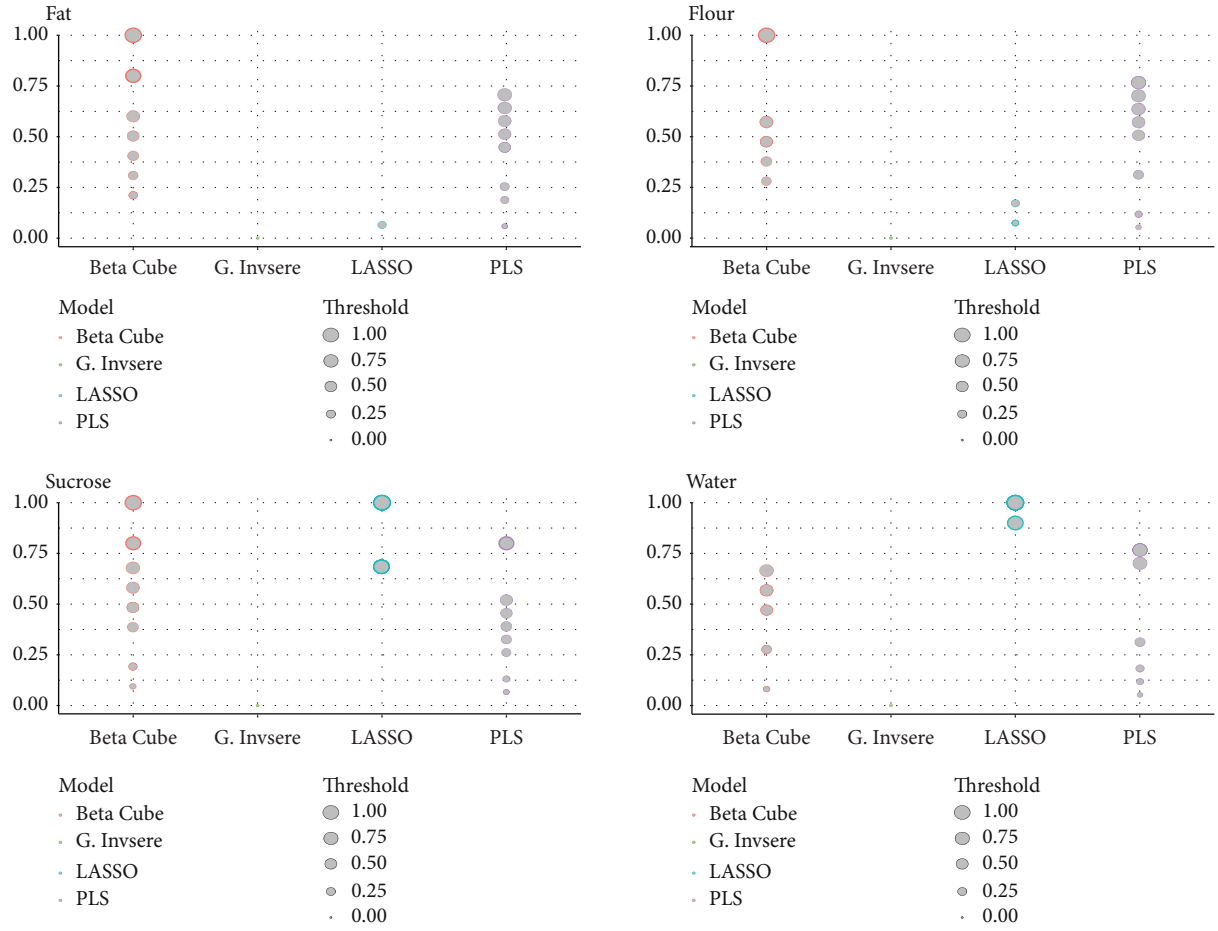

Figure 4: The distribution of threshold and components of biscuit dough is presented for each prediction method.

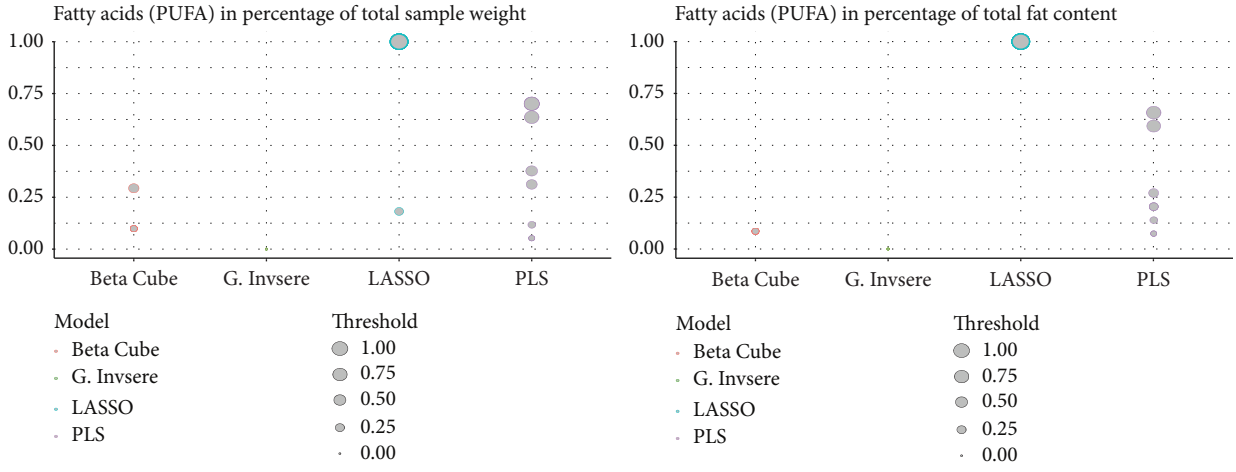

Figure 5: For each prediction approach, the distribution of fatty acid thresholds as a percentage of total sample weight and total fat content is shown.

threshold = 0, the models are including all the $\widehat{\beta}$ coefficients. Beta Cube is picking penalty parameters 0.1 and 0.2. In PLS, the models are considering optimal components in between the range of 0.1 to 0.75. For these data sets, LASSO is not performing well even at the threshold value of 1 (Figure 6).

6. Limitations

Only the used data set makes the suggested results valid. Furthermore, the suggested approach is not robust and cannot be guaranteed to perform well in the presence of outliers.

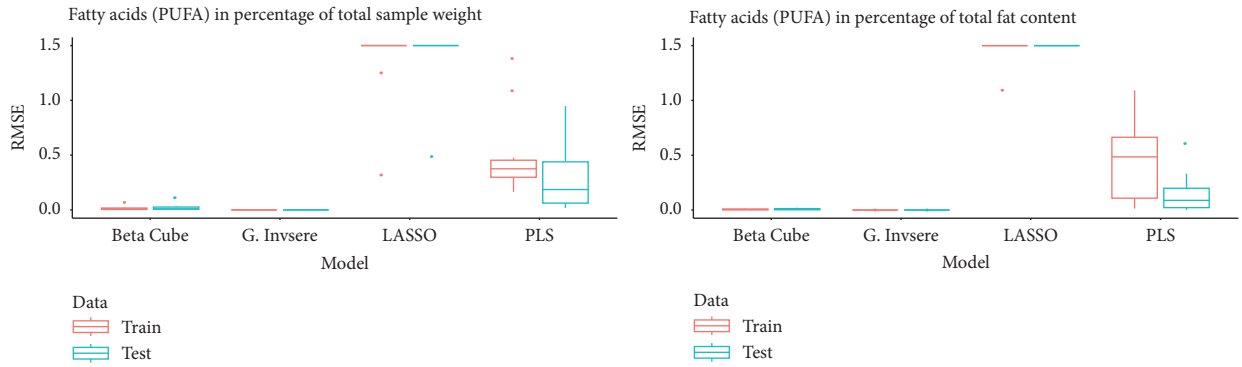

FIGURE 6: The distribution of RMSE of fatty acids as a percentage of total sample weight and total fat content is presented for each prediction method.

7. Conclusions

In terms of high-dimensional chemometric data set prediction capabilities, the presented approaches outperform. Among the data sets modelled are NIR of biscuit dough and Raman of polyunsaturated fatty acids. In addition, the suggested extended inverse and Beta Cube regression techniques yield more consistent results. These regression methods are designed for data sets with a large number of independent variables relative to the sample size. Furthermore, they are effective for multicollinear data sets and identification difficulties. Future studies should delve into the extra features and many uses of the approaches presented.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Rimal, T. Almøy, and S. Sæbø, "Comparison of multiresponse prediction methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 190, pp. 10–21, 2019.

[2] O. Abbas, A. Pissard, and V. Baeten, "3 Near-infrared, mid-infrared, and Raman spectroscopy," in *Chemical Analysis of Food*, pp. 77–134, Elsevier, Amsterdam, Netherlands, 2020.

[3] N. K. Afseth, J. P. Wold, and V. H. Segtnan, "The potential of Raman spectroscopy for characterisation of the fatty acid unsaturation of salmon," *Analytica Chimica Acta*, vol. 572, no. 1, pp. 85–92, 2006.

[4] P. R. Griffiths, "Introduction to the theory and instrumentation for vibrational spectroscopy," *Applications Of Vibrational Spectroscopy In Food Science*, vol. 1, pp. 31–46, 2006.

[5] R. Kizil and J. Irudayaraj, "Raman spectroscopy," *Process analytical technology for the food industry*, vol. 5, pp. 103–134, 2014.

[6] N. K. Afseth, V. H. Segtnan, B. J. Marquardt, and J. P. Wold, "Raman and near-infrared spectroscopy for quantification of fat composition in a complex food model system," *Applied Spectroscopy*, vol. 59, no. 11, pp. 1324–1332, 2005.

[7] D. W. Marquardt and R. D. Snee, "Ridge regression in practice," *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.

[8] J. Ranstam and J. A. Cook, "Lasso regression," *British Journal of Surgery*, vol. 105, no. 10, p. 1348, 2018.

[9] V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang, *Handbook of partial least squares*, Vol. 201, Springer, , Berlin, Germany, 2010.

[10] R. Penrose, "A generalized inverse for matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, pp. 406–413, 1955.

[11] Z. Y. Algamal, M. H. Lee, and A. M. Al-Fakih, "High-dimensional quantitative structure-activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression," *Journal of Chemometrics*, vol. 30, no. 2, pp. 50–57, 2016.

[12] Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih, and M. Aziz, "High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty," *Journal of Chemometrics*, vol. 31, no. 6, p. e2889, 2017.

[13] Z. Y. Algamal, M. H. Lee, A. M. Al-Fakih, and M. Aziz, "High-dimensional QSAR modelling using penalized linear regression model with l 12-norm," *SAR and QSAR in Environmental Research*, vol. 27, no. 9, pp. 703–719, 2016.

[14] U. Indahl, "A twist to partial least squares regression," *Journal of Chemometrics*, vol. 19, no. 1, pp. 32–44, 2005.

[15] Y. Sun, Y. Wang, J. Huang et al., "Quality assessment of instant green tea using portable nir spectrometer," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 240, Article ID 118576, 2020.

[16] T. Næs, O. Tomic, N. K. Afseth, V. Segtnan, and I. Måge, "Multi-block regression based on combinations of orthogonalisation, pls-regression and canonical correlation analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 32–42, 2013.

[17] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 4, pp. 1307–1330, 1986.

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

[19] F. Emmert-Streib and M. Dehmer, "High-dimensionallasso-based computational regression models: regularization,

shrinkage, and selection," *Machine Learning and Knowledge Extraction*, vol. 1, pp. 359–383, 2019.

[20] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistical Learning with Sparsity: The Lasso and Generalizations," *Monographs on statistics and applied probability*, vol. 143, p. 143, 2015.

[21] N. Gauraha, "Introduction to the lasso," *Resonance*, vol. 23, no. 4, pp. 439–464, 2018.

[22] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.

[23] C. R. Rao, "Analysis of dispersion for multiply classified data with unequal numbers in cells," *Sankhya: The Indian Journal of Statistics*, vol. 15, pp. 253–280, 1955.

[24] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, Boca Raton, FL, USA, 2019.

[25] M. Amini and M. Roozbeh, "Optimal partial ridge estimation in restricted semiparametric regression models," *Journal of Multivariate Analysis*, vol. 136, pp. 26–40, 2015.

[26] M. Roozbeh, "Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion," *Computational Statistics and Data Analysis*, vol. 117, pp. 45–61, 2018.

[27] A. Alin, "Multicollinearity. Wiley Interdisciplinary Review," *Wiley Interdisciplinary Review*, vol. 2, no. 3, pp. 370–374, 2010.

[28] W. N. van Wieringen, "Lecture notes on ridge regression," 2015, https://arxiv.org/abs/1509.09169.

[29] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev, "Why the Monte Carlo method is so important today," *WIREs Computational Statistics*, vol. 6, pp. 386–392, 2014.

[30] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.