© 2018 IJSRSET | Volume 4 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Partial Least Squares in Constructing Candidates Model Averaging

Muhammad Arna Ramadhan, Bagus Sartono, Anang Kurnia
Department of Statistics, Bogor Agricultural University, Bogor, Indonesia

ABSTRACT

Model averaging has been developed as an alternative method in regression analysis when number of observations is smaller than number of explanatory variables (also known as high-dimensional regression). Main concept about this method is–weighted average of several candidate models, in order to improve prediction accuracy. There are two steps in model averaging: construct several candidate models and determine weights for candidate models. Our research proposed partial least squares model averaging (PLSMA) as an approach to construct candidate models, while partial least squares (PLS) method was applied during that process to reduce and transform original explanatory variables become new variables that called components. The evaluation of PLSMA is conducted by measured Root Mean Squared Error of Prediction (RMSEP) with simulation data. Compared to other methods, PLSMA has given the smallest RMSEP, so this result indicates that this method had yielded more accurate prediction than other existing methods.

Keywords: Model Averaging, Partial Least Squares, High-Dimensional Regression

I. INTRODUCTION

One problem that often found in regression analysis is when number of observations is smaller than number of explanatory variables. This condition leads to multicollinearity among matrix **X**, which also leads to singular (non-invertible) matrix $\mathbf{X}'\mathbf{X}$ in multiple linear regression. In that case, Ordinary Least Squares (OLS) solution becomes non-unique and leads to poor prediction performance [1].

Model averaging is one of such methods that use weighted average of several models in order to improve prediction accuracy in high-dimensional regression through reduce bias and prediction variance [2]. There are two steps in model averaging. The first step is construct candidate models, which by Perrone [3] are taken by choosing the explanatory variables randomly, or which by Ando and Li [4] based on marginal correlation between the explanatory variable and response variable. The second step is determine weights of candidate models, which are some weights can be chosen based on information criterion (AIC and BIC), Mallows' criterion, Cross-Validation criterion, and weights based on Unbiased Estimator of Risk [5].

Model averaging has been applied in many areas. In economics, Moral-Benito [6] used model averaging to examine deterrent effect of capital punishment. Salaki et al. [7] applied model averaging on experimental design. In genetics, model averaging was applied by Rahardiantoro et al. [8] for predicted the exposure to aflatoxin B_{1}, as well as Posada and Buckley [9] applied this method in phylogenetics.

Our research tries to give a contribution to construct candidate models in model averaging. We propose partial least squares model averaging (PLSMA) as method to construct candidate models. On that process, we apply partial least squares (PLS) to reduce and transform original explanatory variables become new variables called components that satisfy three conditions: highly correlated with response variable, they are have much of the variance among the explanatory variables, and uncorrelated with each other [10].

Our method would be evaluated and compared to the other methods using simulation study. We evaluated these methods by measured root mean squared error of prediction (RMSEP) and correlation between response variable and prediction. The best method is shown by the smallest RMSEP and high-positive correlation that indicate the highest prediction accuracy.

The reminder of the paper is organized as follows. Section 2 describes model averaging, procedure of construction candidate models and selection the weight. Section 3 presents our methods and the algorithm. We presents simulation study to show the merits of the proposed method in Section 4 and conclusions are given in Section 5.

II. MODEL AVERAGING

Model averaging is a weighted average of several models that developed to improve prediction accuracy. Suppose $\mathbf{y}$ is $n \times 1$ response variable vector and $\mathbf{X}$ is explanatory variables of dimension $n \times p$. We denote a set of K candidate models, so model averaging can be written as the function below

$$\mathbf{y} = \sum_{k=1}^{K} w_k M_k \qquad (1)$$

where $\sum_{k=1}^{K} w_k = 1$ and candidate models M_k can be expressed as

$$M_k : \mathbf{y} = \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\varepsilon} \qquad (2)$$

with $\mathbf{X}_k$ is set of explanatory variables to be included in model M_k of dimension $n \times m$ where $(m < n)$ and $\boldsymbol{\beta}_k$ is regression coefficients vector and $\boldsymbol{\varepsilon}$ is random effect vector.

The regression coefficients $\boldsymbol{\beta}_k$ are estimated by ordinary least squares method (OLS):

$$\widehat{\boldsymbol{\beta}}_k = \text{argmin} \, \|\mathbf{y} - \mathbf{X}_k \boldsymbol{\beta}_k\|^2$$

which leads to $\widehat{\boldsymbol{\beta}}_k = (\mathbf{X}_k{}'\mathbf{X}_k)^{-1}\mathbf{X}_k{}'\mathbf{y}$ and least-square prediction $\widehat{\mathbf{y}}_k = \widehat{\mathbf{X}}_k \widehat{\boldsymbol{\beta}}_k$ for $k = 1,2,\dots,K$. Then, prediction of model averaging defined as

$$\widehat{\mathbf{y}} = \sum_{k=1}^{K} w_k \widehat{\mathbf{X}}_k \widehat{\boldsymbol{\beta}}_k \qquad (3)$$

$$\widehat{\mathbf{y}} = \sum_{k=1}^{K} w_k \widehat{\mathbf{y}}_k. \qquad (4)$$

Construction Candidate Models

This section presents two methods for constructing candidate models. The first method was proposed by Perrone (1993). Candidate models are constructed by randomly partition of explanatory variables, then we called RMA (randomized model averaging). Suppose a set of explanatory variables $\mathbf{X} = \{\mathbf{x}_1 \, \mathbf{x}_1 \dots \mathbf{x}_p\}$. By randomly partition, each candidate model M_k contains $\mathbf{X_k} = \{\mathbf{x}_{k1} \, \mathbf{x}_{k2} \dots \mathbf{x}_{km}\}$ for $k = 1,2,\dots K$ and $M_i \neq M_j$ with $i, j \in k$.

The second method is proposed by Ando and Li (2014). They constructed candidate models using ordered explanatory variables which ordered by its marginal correlation with response variable, so explanatory variables in each candidate model have similar correlation with response variables. We called this method as CMA (correlation model averaging). Let $\mathbf{X}^* = \{\mathbf{x}_{[1]} \, \mathbf{x}_{[2]} \dots \mathbf{x}_{[p]}\}$ is the ordered explanatory variables then candidate model M_k contains $\mathbf{X_k} = \{\mathbf{x}_{[(k-1)m+1]} \, \mathbf{x}_{[(k-1)m+2]} \cdots \mathbf{x}_{[km]}\}$.

Determination Weights of Candidate Models

The second step of model averaging is determine weights of candidate models. The weights are determined after construction candidate models. We denote weight of candidate model as w_k and $\sum_{k=1}^{K} w_k = 1$. The higher weight is given to the better candidate model.

There some weight choices for model averaging. Wang et al. (2009) gives some weight choices, there are based on the information criterion (AIC and BIC), Mallow's criterion, Cross-Validation criterion, and weight based on the Unbiased Estimator of Risk.

In this paper, we use AIC weights for model averaging. AIC weights defined as

$$\mathrm{w_k} = \frac{\exp(\mathrm{AIC_k}/2)}{\sum_{k=1}^{K}\exp(\mathrm{AIC_k}/2)} \qquad (5)$$

with

$$\mathrm{AIC_k} = -2\log(\mathrm{L_k}) + 2\mathrm{p} \qquad (6)$$

where $\mathrm{L_k}$ is the maximized likelihood function under k-candidate model and p is the number of parameters.

III. PARTIAL LEAST SQUARES MODEL AVERAGING

Our proposed method for constructing candidate models is named partial least square model averaging (PLSMA). In constructing candidate models, we apply partial least squares (PLS) to reduce and transform original explanatory variables become new variables called components, then these components are used to construct candidate models. We choose PLS in process of constructing candidate models at least for two reasons. First, PLS was developed to handle regression analysis in high-dimensional data (number of observations is smaller than number of explanatory variables) and second, the components that constructed from PLS satisfied three conditions: highly correlated with the response variables, have much of the variance among the explanatory variables, and uncorrelated with each other.

There are three main steps in PLSMA method. First step, half of data is used to estimate the parameters for each candidate model. In the second step, determine AIC weights for each candidate model. The third step, the remaining half of data is used to predict the response variables based on the fitted models, and then the predictions of each candidate model are combined to get the final predictions. Assume that $\boldsymbol{X}$ is $n \times p$ matrix of the explanatory variables and $\boldsymbol{y}$ is vector of the response variable.

For simplicity, the PLSMA algorithm as follows:

Step 1. Split the data into two parts, $\mathrm{Z}^{(1)} = (\mathbf{X}_\mathrm{i}, \mathbf{y}_\mathrm{i}), 1 \leq \mathrm{i} \leq 2\mathrm{n}/3$, and $\mathrm{Z}^{(2)} = (\mathbf{X}_\mathrm{i}, \mathbf{y}_\mathrm{i}), 2\mathrm{n}/3 + 1 \leq \mathrm{i} \leq \mathrm{n}$.

Step 2. Resampling 75% observation of data $\mathrm{Z}^{(1)}$ and do the partial least squares (PLS) process to get the x-loadings $\mathbf{h}$, this is used to reduce dimension of $\mathbf{X}$ and transform $\mathbf{X}$ into new variables, components $\mathbf{T} = \mathbf{Xh}$. Note that $\mathbf{h}$ is $\mathrm{p} \times \mathrm{m}$ matrix of x-loadings, where m is the number of components in each candidate model ($\mathrm{m} < \mathrm{p}$).

Step 3. Construct candidate model by regressing $\mathbf{T}$ onto $\mathbf{y}$.

Step 4. Repeat step 2 and step 3 until k candidate models are constructed.

Step 5. Compute the weights for each candidate model

$$\mathrm{w_k} = \frac{\exp(\mathrm{AIC_k}/2)}{\sum_{k=1}^{K}\exp(\mathrm{AIC_k}/2)}$$

Step 6. Compute the prediction of each candidate model using the remaining half of the data $\mathrm{Z}^{(2)}$.

Step 7. Let

$$\hat{\mathbf{y}} = \sum_{k=1}^{K} \mathrm{w_k}\hat{\mathbf{y}}_\mathrm{k}$$

be the final PLSMA prediction.

IV. SIMULATION STUDY

Data Simulation

Our proposed method, PLSMA is evaluated using data simulation. We adopted the settings of Ando and Li (2014) with few modifications. We determine the number of observations $\mathrm{n} = 100$ and generate the explanatory variables $\mathrm{p} = 2000$ from multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathrm{S} = \mathrm{s_{ij}}$ with $\mathrm{s_{ij}} = \rho^{|\mathrm{i}-\mathrm{j}|}$ and set $\rho = 0.8$. The response variable is generated through the regression model with intercept 100 and regression coefficient 1. We set the significant regression coefficient $s = 50$ and be spaced evenly $i = 40(j - 1) + 1, j = 1,2, \dots ,50$. Random effects are generated from normal distribution with mean 0 and standard deviation 4.

V. RESULT AND DISCUSSION

In this simulation study, we set number of candidate models $k = 20$ and some number of regressors in candidate model $m = \{5,10,15,20\}$. These settings are applied in our method and compared to other methods. RMSEP (root mean squared error of prediction) is used as an indicator to measure the prediction accuracy. The performance measure RMSEP after 50 simulation runs shown in Figure 1. These results showed that our proposed method produced the smallest RMSEP in each condition of m, so could be indicated that our method gives the better performance than the other methods.

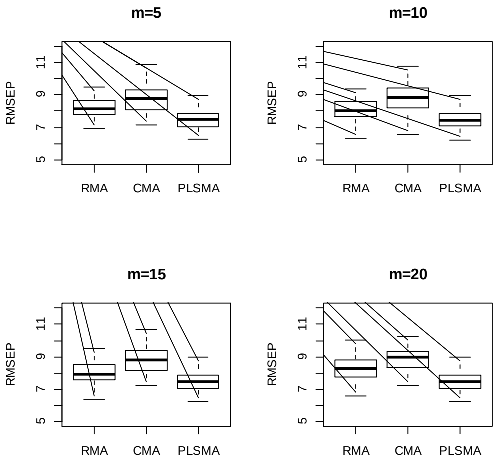

Figure 1. Boxplots of the performance measure RMSEP

We showed plots of actual value of the response variable and prediction in Figure 2. Both of methods, RMA and CMA yielded random pattern, while PLSMA yielded linear pattern. This pattern showed that PLSMA produced higher correlation between actual value and prediction than RMA and CMA. For details, we also showed correlations that produced by each method in Figure 3.

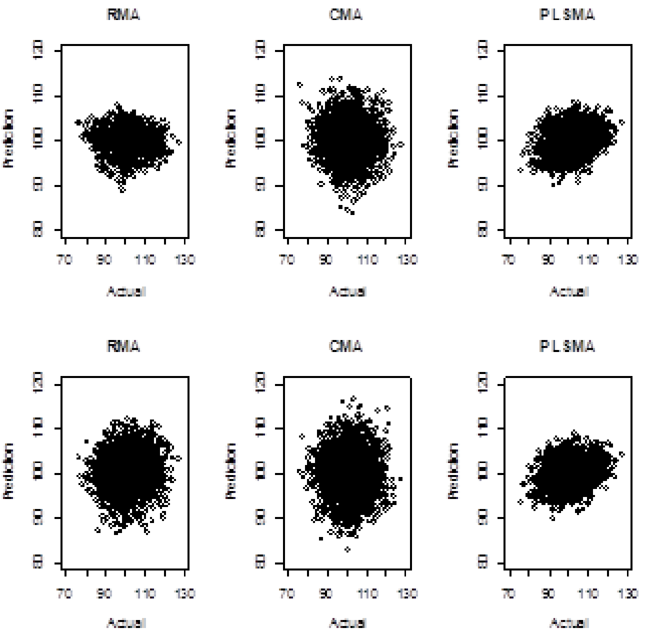

Figure 2. Plot actual and prediction

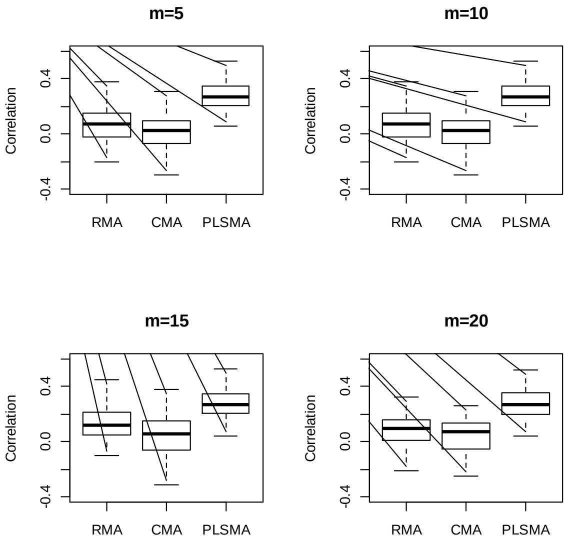

Figure 3. Boxplot of correlation between actual and prediction

In Figure 3. we showed boxplots of correlation value between actual and prediction for each method. PLSMA produced the highest correlation than RMA and CMA for eac m. The correlation produced by PLSMA also showed that this method increases the prediction accuracy. As an additional, we showed that PLSMA has more homogen residual than RMA and CMA. We showed it in plots of prediction and residual in Figure 4.

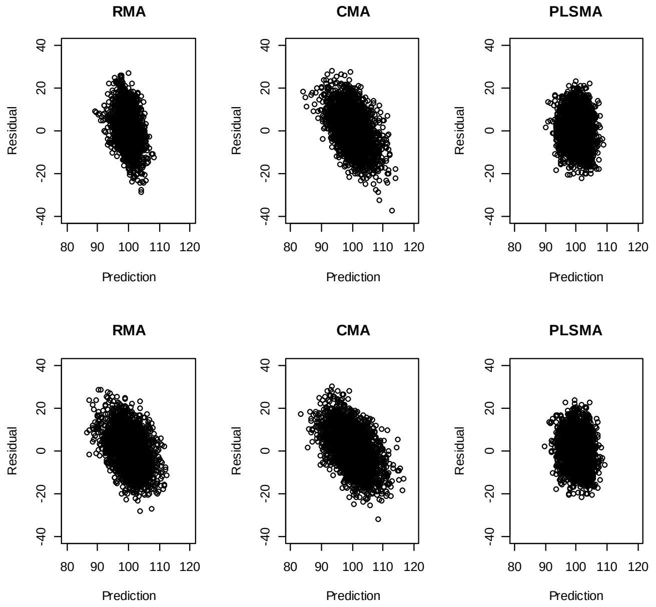

Figure 4. Plot Prediction and Residual

The simulation study had been shown that PLSMA produced the highest prediction accuracy than the other methods.

VI. CONCLUSION

Model averaging could be better choice in regression analysis when number of observations is much smaller the number of explanatory variables. Our method, PLSMA was developed to construct candidate models in model averaging. The simulation study had been shown that PLSMA produced more accurate prediction, in terms of some indicators evaluated such as RMSEP, correlation value of actual and prediction, and homogenity of residuals.

VII. REFERENCES

[1]. Naes T., Isakson T., Fearn T., and Davies T. 2002. A user-friendly guide to multivariate calibration and classification, NIR Publications, Chicester.

[2]. Hansen, B. E. 2007. Least squares model averaging. Econometrica, 75(4), 1175-1189.

[3]. Perrone, M. P. 1993. Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization (Doctoral dissertation, Brown University).

[4]. Ando, T. and Li, K. C. 2014. A model-averaging approach for high-dimensional regression. Journal of the American Statistical Association, 109(505), 254-265.

[5]. Wang, H., Zhang, X., and Zou, G. 2009. Frequentist model averaging estimation: a review. Journal of Systems Science and Complexity, 22(4), 732.

[6]. Moral-Benito, E. 2015. Model averaging in economics: An overview. Journal of Economic Surveys, 29(1), 46-75.

[7]. Salaki, D. T., Kurnia, A., and Sartono, B. 2016. Model averaging method for supersaturated experimental design. In IOP Conference Series: Earth and Environmental Science (vol. 31, no. 1, p. 012016). IOP Publishing.

[8]. Rahardiantoro, S., Sartono, B., and Kurnia, A. 2017, March. Model averaging for predicting the exposure to aflatoxin B1 using DNA methylation in white blood cells of infants. In IOP Conference Series: Earth and Environmental Science: (vol. 58, no. 1, p. 012019). IOP Publishing.

[9]. Posada, D., and Buckley, T. R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic biology, 53(5), 793-808.

[10]. Dunn, W. J., Scott, D. R., and Glen, W. G. 1989. Principal components analysis and partial least squares regression. Tetrahedron Computer Methodology, 2(6), 349-376.