© 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Model Averaging Approach in Calibration Model

Deiby Tineke Salaki*[1], Anang Kurnia[2], Arief Gusnanto[3], I Wayan Mangku[4], Bagus Sartono[2]

[1]Department of Mathematics, Sam Ratulangi University, Manado, North Sulawesi, Indonesia

[2]Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia

[3]Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom

[4]Department of Mathematics, Bogor Agricultural University, Bogor, West Java, Indonesia

ABSTRACT

This article deals with model averaging as an alternative regression technique for high-dimensional data especially in chemometrics where statistical approach is used to extract any information contained in a chemical dataset. Our simulation study indicated that model-averaging (MA) works better in high-correlated data than in low-correlated data. The result also designated MA with weighting procedure based on Mallows' *Cp* and Jackknife criteria produce better predictions compared to Akaike information criterion (AIC)-based of weight if the candidate models are constructed by randomly grouping the covariates. Moreover, the prediction performance tent to increase along with the number of variables in a candidate model. We illustrated the methods to regress the concentration of curcuminoid in curcumin specimen as a function of their spectra determined by Fourier Transform Infra-red (FTIR) instrument.

Keywords: AIC, Calibration model, Curcumoid, FTIR, High-dimensional data, Jackknife ,Mallows Cp, Model averaging.

I. INTRODUCTION

A dataset resulted from spectroscopy such as Near Infra Red (NIR) and Fourier Transform Infra-Red (FTIR) is a kind of chemical dataset which represents the quantity of heat absorbed or emitted by certain substance over several wave lengths. Its commonly used to quantify concentration of a certain substance containing in a sample.

Calibration model aims at building a functional relationship between concentration of interest component in certain specimen and a huge number of absorbencies which is much bigger than the number of observations. In this condition, the classical regression approach such as ordinary least squares (OLS) is no longer suitable to be implemented due to the presence of high multi-collinearity among the explanatory variables.

In calibration models some well-known approaches have been proposed so far including continum regression [1] and Bayesian approach [2]. Sparse alternatives using random effect approach for calibration model has also been studied by Gusnanto and Pawitan [3].

This article considered an alternative approach where the final model relies on a weighted average of a set of approximation models known as model averaging (MA). There are two well-known points of view in MA namely Bayesian MA (BMA) and Frequentist MA (FMA). While the former computes posterior probabilities for each of candidate models and use them as weights [4], the latter does it based on certain criteria such as Akaike Information Criteria (AIC) [5], Mallows [6] and Jackknife [7]. Application of MA in calibration data which has not popular yet. In other bioinformatic problems, however, it has been widely appeared in many fields of interest such as in genetic problem by Rahardiantoro et al. [8], supersaturated experimental design by Salaki et al. [9] and public health by Ando and Li [10].

Mallows model-averaging (MMA) and Jackknife model-averaging (JMA) are least squares model-averaging approaches in which the construction of candidate models use nested modelling. In application to high-dimensional data such as calibration model the setting is impossible to be applied. Factually, computational complexity will rapidly increase as the number of predictors becomes large. Furthermore, in order to apply the JMA to high-dimensional data, Ando and Li [10] modified JMA in way of preparing candidate models by taking marginal correlation between independent variable and response into account.

In this study, we prepare the candidate models by grouping randomly the variables into certain number of subsets. This mode has been conducted previously for instance by Ramadhan *et al.* [11] with AIC as the weight criterion in a simulation study. We use the same approach in comparing with Mallows and Jackknife weight in a simulation study before applying the best of them in a real dataset. Thus, we employ MA for building calibration model of a dataset from FTIR to make comparison of AIC MA (AMA), MMA and JMA performances.

The rest of this article is organized as follow. Section II discusses about methods and material. The next Section III details results and discussion and the last section is used for some conclusion statements.

II. METHODS AND MATERIAL

A. Model Averaging

Let $\mathbf{X}$ be a matrix of spectra determined by FTIR instrument with p wavelengths (variables) and n observation. The matrix size is $n \times p$ with $p \gg n$. The intercept term is dropped by centering each variable to have zero mean. Suppose $\mathbf{y}$ be an n-vector of curcuminoid concentration detected by High Performance Liquid Chromatography served as a response variable. Regression model parameters is represented as a p-vector $\boldsymbol{\beta}$. The model of response $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (1)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2,, \dots, \mathbf{x}_p]$. We assume $\boldsymbol{\epsilon}$ an n-vector of error term following the normal distribution with mean zero and variance σ^2. Least squares (LS) estimate of regression parameters $\boldsymbol{\beta}$ can be written as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\boldsymbol{y} \qquad (2)$$

In case of $p \gg n$, $\mathbf{X}^T\mathbf{X}$ is not invertible that LS methods cannot be implemented.

While selection model chooses a single best model from several competing ones as the final result, model averaging employs all of them by weighted averaging those models to tackle uncertainty model containing in the prior approach [12]. There are three important aspects in the approach, i.e. construction and estimation candidate models and criteria of weight selection, as described bellow.

Construction and Estimation of Candidate Models

This section describes development and estimation of candidate models. Following the settings of Ando and Li [10], we rewrite the equation (1) as

$$\mathbf{y} = \textstyle\sum_{j=1}^{p} \beta_j \mathbf{x}_j + \boldsymbol{\epsilon}. \qquad (3)$$

The number of variables in a subset is limited to a value that is less than the number of observations. In order to develop K candidate models we separate randomly the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}$ into K subsets. For simplicity, those subsets are set to have the same cardinal number, that is $\frac{p}{K}$ elements. As we employed least squares as the estimation methods, we restricted $\frac{p}{K} \leq n$. Thus, we have a sequence of candidate linear models $M_1, M_2, \dots, M_K$ where the candidate model M_K can be stated as:

$$M_k\colon \mathbf{y} = \mathbf{X}_{(k)}\boldsymbol{\beta}_{(k)} + \boldsymbol{\epsilon} \qquad (4)$$

Where $\mathbf{y} = (y_1, y_2, \dots, y_p)^T$; $\mathbf{X}_{(k)}$ is a design matrix of model M_k with size $n \times \frac{p}{K}$; $\boldsymbol{\beta}_{(k)}$ is a $\frac{p}{K}$ -vector of parameter associated with $\mathbf{X}_{(k)}$ and $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(\epsilon_1, \epsilon_2, \dots, \epsilon_p)^T$.

If $\widehat{\boldsymbol{\mu}}_{(k)}$ is the LS prediction of $\mathbf{y}$ based on model M_k then MA prediction $\widehat{\boldsymbol{\mu}}_{MA}$ of equation (3) can be formulated as:

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_{MA} &= \textstyle\sum_{k=1}^{K} w_k \widehat{\boldsymbol{\mu}}_{(k)} \\ &= \textstyle\sum_{k=1}^{K} w_k [\mathbf{X}_{(k)}(\mathbf{X}_{(k)}^T \mathbf{X}_{(k)})^{-1} \mathbf{X}_{(k)}^T \mathbf{y}] \\ &= \textstyle\sum_{k=1}^{K} w_k \boldsymbol{H_k} \boldsymbol{y} = \boldsymbol{H}(\boldsymbol{w})\boldsymbol{y} \qquad (5)\end{aligned}$$

where $\boldsymbol{H}(\boldsymbol{w}) = \sum_{k=1}^{K} w_k \boldsymbol{H_k}$ is the corresponding hat matrix and $\boldsymbol{w}$ is a k-vector of weight such that $\sum_{k=1}^{K} w_k = 1$ and w_k is weight term corresponding to candidate model M_k.

Weight Selection Criteria

In model averaging where all competing models are considered to contribute to final model, weight criteria play important role to determine the prediction performance. AIC criterion was firstly introduced by Burnham & Anderson [5] which is based on AIC score. For a candidate model M_k with $d(M_k)$ length of parameter vector and log likelihood function L_k, the weight w_{k} is defined as

$$w_{\mathrm{k}} = \frac{\exp(-\Delta_{\mathrm{k}}/2)}{\sum_{k=1}^{K} \exp(-\Delta_{\mathrm{k}}/2)} \qquad (6)$$

Here, $\Delta_k = \mathrm{AIC}_k - \mathrm{AIC}_{min}$ is a measure of candidate model M_k relative to the best model that is model with minimum AIC (AIC_{min}). Consequently, the better a candidate model the bigger the weight to be assigned in it.

The MMA which is proposed by Hansen [6] select the weight of averaging by minimizing a Mallow Cp criterion

$$\mathrm{C_n}(\boldsymbol{w}) = \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_{MA}\|^2 + 2\hat{\sigma}^2 \boldsymbol{w}^T \Phi \qquad (7)$$

where $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_K)^T$; φ_k denotes the number of covariates in candidate model M_k. The selected weight is the weight vector $\widehat{\boldsymbol{w}}$ that minimizes the criteria (7), that is

$$\widehat{\boldsymbol{w}} = (\widehat{w}_1, \widehat{w}_2, \dots, \widehat{w}_K)^T = \arg\min_{w \in Q} \mathrm{C_n}(\boldsymbol{w})$$

where $= \{\boldsymbol{w} \in [0,1]^K;\ \sum_{k=1}^{K} w_k = 1\}$.

Jackknife criteria of weight selects the weight using the leave-one-out cross-validation (LOOCV) or Jackknife as previously used in Hansen and Racine [7].

Following their setting, we define $\widetilde{\boldsymbol{\mu}}_k = (\tilde{\mu}_k^{-1}, \tilde{\mu}_k^{-2}, \dots, \tilde{\mu}_k^{-n})^T$ as an n -vector. Here, $\tilde{\mu}_k^{-\alpha}$ denotes the predicted value of the αth observation resulted from a training dataset developed by deleting the αth observation $(y_\alpha, \boldsymbol{x}_\alpha)$ and based on model M_k. Referring to the hat matrix H_k in equation (5), we define $\widetilde{H}_k = D_k(H_k - I) + I$; D_k is an n -diagonal matrix where the α th diagonal element equals to $(1 - h_{k\alpha})^{-1}$; $h_{k\alpha}$ is the α th diagonal element of H_k. The leave-one out predictor is formulated as

$$\widetilde{\boldsymbol{\mu}} = \textstyle\sum_{k=1}^{K} w_k \widetilde{\boldsymbol{\mu}}_k = \sum_{k=1}^{K} w_k \widetilde{H}_k \mathbf{y} = \widetilde{H}(\boldsymbol{w})\mathbf{y}$$

where $\widetilde{H}(\boldsymbol{w}) = \sum_{k=1}^{K} w_k \widetilde{H}_k$.

The sum of squared residuals of leave-one-out predictor is used to form the cross-validation criterion $\mathrm{CV}(\boldsymbol{w}) = \|\boldsymbol{y} - \widetilde{\boldsymbol{\mu}}\|^2 = \left\|\boldsymbol{y} - \widetilde{H}(\boldsymbol{w})\mathbf{y}\right\|^2$ (8)

The selected weight is the weight vector $\widehat{\boldsymbol{w}}$ that minimizes the criteria (8), that is

$$\widehat{\boldsymbol{w}} = (\widehat{w}_1, \widehat{w}_2, \dots, \widehat{w}_K)^T = \arg\min_{w \in Q} \mathrm{CV}(\boldsymbol{w}).$$

B. Data

In this section we explain about the simulated data and a real spectroscopy dataset to build a calibration model.

Simulated Data

As the spectroscopy dataset is commonly a high-dimensional data with high correlation between its covariates, we generate a dataset containing the sample size n= 300 and p=1000 predictors. As much as 200 predictors having index $i = 10(j-1)+1; j = 1,\dots,200$ are set as the true predictors by setting $\beta_i = 1$. The covariance matrix is defined as $C = [c_{ij}]$ where

$$c_{ij} = \begin{cases} 0.95\,; i \neq j \\ 1 \quad ; i = i \end{cases}.$$

In order to investigate how the prediction performance behaves, we varied the error terms generated from normal with mean 0 and variance 0.1, 0.3 and 0.5.

Spectroscopy Dataset

In this article, we employ a spectroscopy dataset of curcuminoid active compound in curcuma specimen obtained from the observation of curcuma herbs. The dataset is produced by the Post Graduate Research Team 2003-2005 which is a collaboration between the Statistics Department with Biopharmaca Study Centre Bogor Agricultural University, Indonesia as previously described in [1] and [2]. The dataset is resulted from calibration of FTIR instrument with spectra from curcuma and comprised of 20 observations measured in 1866 different wavelengths. It consequently creates a covariate matrix of 20 rows and 1866 columns.

III. RESULTS AND DISCUSSION

This section is dedicated to present the result of both simulation study and calibration model of curcumin dataset to compare the prediction performance of AMA, MMA and JMA.

A. Simulation Study

In order to bring out the performance of model averaging by using the three different weight criteria, we evaluate their prediction in term of root mean squared error prediction (RMSEP). We vary the number of covariates in a candidate model in several value to investigate, how the quantity influences the performance of MA.

We employ 5-fold cross validation in this study. After randomly rearranging the order of observation, we split the data into 5 equally part each containing 300/5=60 observations. Thus, when the observations in the single i -fold is deleted-out to serve as validation dataset, the remain folds are used for the training data to build a model to be evaluated by using the corresponding validation data.

The step-by-step of model averaging can be itemized as below:

Step 1. Prepare some candidate models by randomly dividing all the independent variables in training set into K groups for K=100, 50, 25, 10 and 5. Consequently, the corresponding number of covariates in a candidate model are varied as nv= 10, 20, 40, 100 and 200. Thus, if the number of covariates in a candidate model is set to 200, we have a sequence of candidate models $M_1,\dots,M_5$. The variables belong to group k are used to build a design matrix for candidate model M_k .

Step 2. Estimate the parameters of each of K candidate models $\hat{\boldsymbol{\beta}}_{(1)}, \hat{\boldsymbol{\beta}}_{(2)}, \dots, \hat{\boldsymbol{\beta}}_{(K)}$ by using least squared method and then compute the corresponding prediction $\hat{\boldsymbol{\mu}}_{(1)}, \hat{\boldsymbol{\mu}}_{(2)}, \dots, \hat{\boldsymbol{\mu}}_{(K)}$ by using the validation dataset.

Step 3. Select weight term w_k for the k-candidate model by using AIC, Mallows and jackknife criteria separately.

Step 4. Computing the prediction of model averaging $\hat{\boldsymbol{\mu}}_{MA} = \sum_{k=1}^{K} w_k \hat{\boldsymbol{\mu}}_{(k)}$.

Step 5. Repeat the step 1 until 4 by using the next fold and so on until fold 5.

Step 6. Compute the RMSEP to represent the prediction performance,

$$\text{RMSEP} = \frac{1}{300}\sum_{i=1}^{300}(y_i - \hat{\mu}_{MA,i})^{\frac{1}{2}}$$

Where y_i and $\hat{\mu}_{MA,i}$ represent the actual and predicted data of the i-th observation respectively

The average of RMSEP of model averaging with different weight criteria are listed in Table I. According to the table, model averaging performance is significantly influenced by the criteria of weight selection as well as the number of covariates in a candidate model. The results in all type of error variance show the same trend, that average of RMSEP of all types of weight increased as the nv gets higher and then decreases at nv=200.

Table 1. Mean Of Rmsep Based On 3 Types Of Ma Over 500 Runs

nv	Model Averaging		
	AMA	MMA	JMA
$\sigma^2 = 0.1$			
10	7.673	2.380	2.334
20	4.959	2.221	2.231
40	4.189	2.212	2.208
100	3.376	2.126	2.126
200	6.428	3.180	3.180
$\sigma^2 = 0.3$			
10	6.265	2.214	2.154
20	4.920	2.164	2.151
40	3.959	2.232	2.231
100	3.326	2.019	2.019
200	5.669	2.853	2.853
$\sigma^2 = 0.5$			
10	7.111	2.543	2.463
20	5.211	2.308	2.309
40	4.401	2.383	2.383
100	3.680	2.295	2.295
200	5.510	3.215	3.215

We can infer that the best performance of a model averaging approach reaches the best when the number of covariates in a candidate model is set to nv=100.

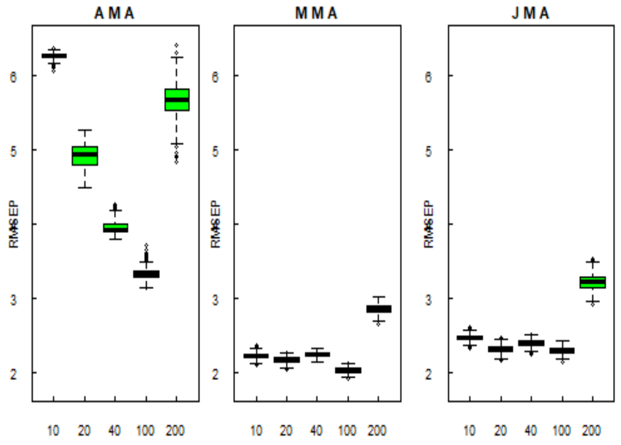

Figure 1. Boxplots of RMSEP of AMA, MMA and JMA.

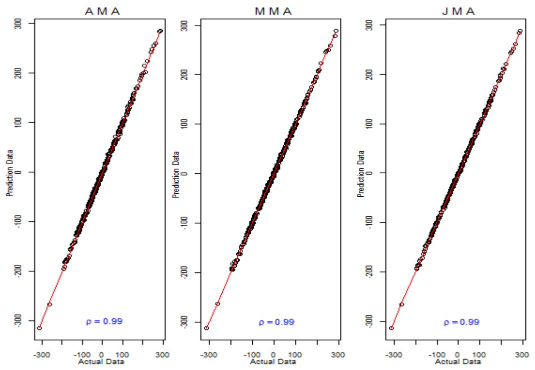

Figure 2. Plot of validation data versus prediction data of the AMA, MMA and JMA

From the Figure 1, we can see the boxplot of RMSEP of the three different weights of MA. The performance resulted from MA have the same trend, however, the variance of RMSEPs stemming from AMA seem higher than the two others. According to the picture we can infer that MMA and JMA are more recommended to be used as the alternative in modelling the high-dimensional data.

Figure 2 shows that the correlation value between validation data and prediction data is significantly support the other remarks. As printed in the figure, the correlation value (ρ) for each method equals to 0.99. The correlation value shows that the trend of predicted data works along with of observed data.

According to Figure 3 where the plot of residuals are randomly scattered and making a relative symmetric band.

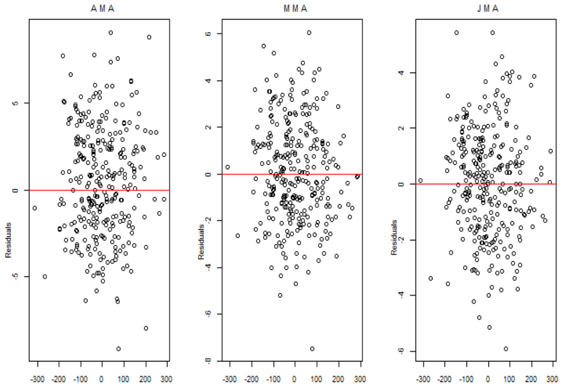

Figure 3. Plots of Residuals versus Prediction Data

B. Calibration Model of Spectroscopy Dataset

Figure 4., displays the spectroscopy data of curcuminoid after being centred on their associated variables means. In this analysis, we set the number of candidate models to 310 each with 6 variables that the OLS method can be implemented. The intercept term is dropped as the data have been firstly centred to have mean zero.

Our result show that RMSEP value resulted from AMA, MMA and JMA equal to 0.748, 0.669 and 0.664 respectively. The results show similar trend with the simulation result that the MMA and JMA outperform AIC weight-based model averaging. In term of RMSEP.

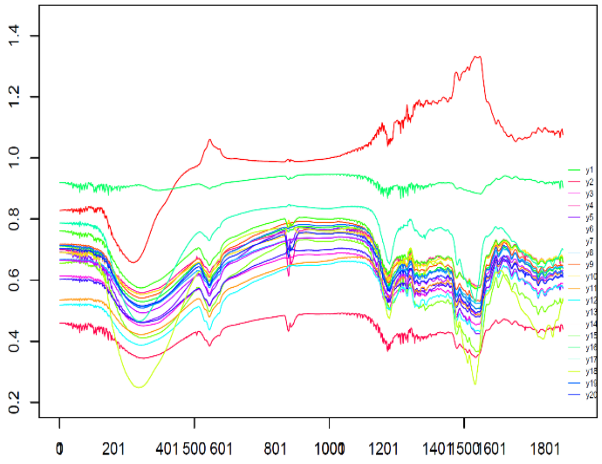

Figure 4. FTIR spectra plot of 20 batches of curcumoid

The correlation between validation versus its corresponding prediction data equal to 0.879 and 0.849 respectively. The higher the correlation value between covariates the better the model performance resulted by model averaging. Moreover, the value shows that trend of observed data works along with its corresponding predicted data and guarantees that the resulted prediction model fits the data.

IV. CONCLUSION

Prediction performance of several weight criteria of model averaging has been compared and applied in calibration model. Based on our simulation study, model averaging with Mallows and Jackknife criteria of weight outperform one with AIC weight. The result also shows that the number of covariate in a candidate model significantly determines the prediction performance. In spectroscopy dataset that suffers from multi-collinearity, model averaging can be employed as an alternative approach to improve the prediction performance in calibration model.

V. REFERENCES

[1]. S. Setiawan, K. A. Notodiputro, Continum regression with discrete wavelet transformation preprocessing, Jurnal ILMU DASAR 8 (2) (2009) 103–109.

[2]. Erfiani, Pengembangan model kalibrasi dengan pendekatan bayes (studi kasus tanaman obat), Ph.D. thesis, Institut Pertanian Bogor, Indonesia (2005).

[3]. A. Gusnanto, Y. Pawitan, Sparse alternatives to ridge regression: a random 205 effects approach, Journal of Applied Statistics 42 (1) (2015) 12–26.

[4]. A. E. Raftery, D. Madigan, J. A. Hoeting, Bayesian model averaging for linear regression models, Journal of the American Statistical Association 92 (437) (1997) 179–191.

[5]. K. P. Burnham, D. R. Anderson, Model selection and multimodel inference: a practical

information-theoretic approach, Springer Science & Business Media, 2003.

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

[7]. B. E. Hansen, J. S. Racine, Jackknife model averaging, Journal of Econometrics 167 (1) (2012) 38–46.

[8]. S. Rahardiantoro, B. Sartono, A. Kurnia, 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, IOP Publishing, 2017, p. 012019.

[9]. D. T. Salaki, A. Kurnia, B. Sartono. Model averaging method for supersaturated experimental design, in: IOP Conference Series: Earth and Environmental Science, Vol. 31, IOP Publishing, 2016, p. 012016.

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

[11]. M.A. Ramadhan, B. Sartono, A. Kurnia. 2018. Partial Least Squares in Constructing Candidates Model Averaging. IJSRSET (4)1: 1459-1463)

[12]. Claeskens G, Hjort N. 2008. Model Selection and Model Averaging. Cambridge University, New York