Forum Statistika dan Komputasi : *Indonesian Journal of Statistics* available online at:
ISSN : 0853-8115 *journal.ipb.ac.id/index.php/statistika*
Vol. 20 No.2, Oktober 2015, p: 116-119

MODEL AVERAGING, AN ALTERNATIVE APPROACH TO MODEL SELECTION IN HIGH DIMENSIONAL DATA ESTIMATION

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

1) Graduate students of Bogor Agricultural University, Indonesia
2) Statistics Department of Bogor Agricultural University, Indonesia
3) Statistics Department of Leeds University, UK
4) Mathmetatics Department of Bogor Agricultural University, Indonesia

ABSTRACT

Model averaging is an alternative approach to classical model selection in model estimation. The model selection such as forward or stepwise regression, use certain criteria in choosing one best model fitted the data such as AIC and BIC. On the other hand, model averaging estimates one model whose parameters determined by weighted averaging the parameter of each approximation models. Instead of conducting inference and prediction only based one best chosen model, model averaging covering model uncertainty problem by including all possible model in determining prediction model. Some of its developments and applications also challenges will be described in this paper. Frequentist model averaging will be preferential described.

Keywords : model selection, frequentist model averaging, high dimensional data

INTRODUCTION

Multiple regression methods plays very important role in data analysis to describe the relationship between a response variable and explanatory ones or predictors. It also deals with prediction of future value of a response and selects which predictors contribute to the model. The main goal of the method is to estimate the best model fitted the data. The challenges in the trading off between descriptive accuracy and parsimony of the chosen model become the motivation of its development.

When population regression is not available, it is replaced by regression over the training set, which sometimes doesn't work well and remain uncertainty. Model selection and model averaging are two popular approaches to deal with model uncertainty appeared in model estimation of high dimensional data. Another approach which is based on penalized least squares, conducted selection variables and estimation simultaneous. Some of them are, for example SCAD (Fan & Li, 2001) and LASSO (Tibhsirani, 1996).

Model selection chooses the one among all candidate models that is regarded as the most accurate description. The further inference and prediction then will be based on the chosen model and surely neglect the rest. The process, however, is complicated by the fact that the more variables included in the model the more accurate in prediction. Meanwhile, model with fewer variables is preferred as its efficiency. The first step of the methods is setting an estimation criterion such as Cp Mallow, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) before selecting from a set of candidate models which scores most highly according to related criteria. Unfortunately, different criteria favor different model to yield a good approximation.

Model averaging is an alternative to model selection. Instead of relying on only one best model, the methods refers both inference and prediction to the average over a set competing model in particular manner. There are two well-known approaches: Bayesian Model Averaging (BMA) and Frequentist Model Averaging (FMA). BMA computes posterior probabilities for each of the models and use them as weights. In real application, the way of setting prior probabilities and how to deal with the priors when they are in conflict is still in debate. On the other hand, FMA requires no prior. The key issues of the methods are for

example, weight selection (Hansen B. , 2014) and the way of construction the candidate model (Ando & Li, 2014).

Another alternative to model selection and model averaging is penalized least squares based which conduct selection model and estimation simultaneously such as Lasso (Least Absolute Shrinkage and Selection Operator) (Tibhsirani, 1996) and SCAD (Smoothly Clipped Absolute Deviation) (Fan & Li, 2001).

Development of model averaging in frequentist perspective has been discussed in a huge literatures such as (Burnham & Anderson, 2002) and (Hansen B. E., 2007). This article presents some recent advancement and challenges of model averaging approach for model estimation in high dimensional data, especially when the number of covariates is highly exceeded of sample. The focus will be more on FMA. Further and detail review of BMA is available in some papers such as (Raftery, Madigan, & Hoeting, 1997) and (Magnus, Powell, & Prufer, 2010).

The rest of the paper is organized as follows. Section 2 discusses the frame work of FMA. Model weights selection developments presents in section 3. Some approaches in construction the model candidate describe in section 4 and then conclusion provides in the last section.

MODEL AVERAGING FRAME WORK

The description of the frame work relies on Ando and Li's (2014). Consider an n x p matrix $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ of covariate vectors with $E(\varepsilon) = 0$ and var $(\varepsilon) = \sigma^2 > 0$ such that $\boldsymbol{\varepsilon}$ is not depend on $\boldsymbol{X}$. Multiple regression model to independent variable y, can be expressed as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

With parameter $\beta = (\beta_1, \beta_2, ., \beta_p)^T \in \mathbb{R}^p$

Because not all the covariates give contribution to prediction model, it can be claimed the existence of a set $T_i = \{j; |\beta_j| > 0\}$ $i = 1,...,M$; $n(T_i) = s$ and $s < p$. It means T_i is the set of covariate index belongs to model M_i. Consequently, the ith model candidate can be written as follow

$$M_i : y_i = \sum_{j \epsilon T_i} \beta_j X_j + \varepsilon \quad (2)$$

According to certain criteria, model selection methods will choose a single best model:

$$\hat{y}_i = \sum_{j \in T_i} \hat{\beta}_j X_j \quad (3)$$

For $i \in \{1,2,...,K\}$. Where $\hat{\beta}_j$; $j \in T_i$ is estimator for β referred to the ith model.

Recall the model (1) above and the set $M = \{M_1,...,M_K\}$; M_i is the same as equation (2). Let w_i is the weight associated with $\hat{\beta}_i$ and satisfied

$$\sum_{i=1}^{K} w_i = 1 \quad (4)$$

and $\hat{\beta}_i$ is parameter vector estimated from model M_i, then *model averaging* estimator of the parameter β in model (1) takes the form

$$\hat{\bar{\beta}} = \sum_{i=1}^{K} w_i \hat{\beta}_i \quad (5)$$

where $\hat{\beta}_i$ is the estimator of β on the basis of the-i candidate model.

Consider the prediction related to model M_i :

$$\boldsymbol{\mu}_i = X_i \hat{\boldsymbol{\beta}}_i \quad (6)$$

where X_i is covariate matrix of model M_i.

Then the MA estimator of μ can be expressed as

$$\hat{\mu}(\boldsymbol{w}) = \sum_{m=i}^{M} w_i \boldsymbol{\mu}_i \quad (7)$$

Model Weights

Based on equations (5) and (7) in the section before, weights of the model determine the MA estimator. This fact becomes motivation in developing criteria in selecting the weight to in getting better estimation.

1. Jacknife Model Averaging (JMA)

If $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)'$, then jackknife model averaging estimator to $\boldsymbol{\mu}$ is

$$\hat{\mu}(W) = \sum_{k=1}^{M} w_k \tilde{P}_k Y \triangleq \tilde{P}(W)\, Y$$

with $\tilde{P}_k = \widetilde{D}_k (P_k - I_n) + I_n$

and $\widetilde{D}_k$ is diagonal matrix diagonal wit n dimension and i-th element can be written as $\left(1 - h_{ii}^k\right)^{-1}$; $h_{ii}^k = X_{k,i}(X_k^t X_k)^{-1} X_{k,i}^t$. $X_{k,i}$ is i-th row of X_k.

$\hat{\varepsilon}_{i(m)}$

$= Y_i - X'_{i(m)} \left(X'_{-i(m)} X_{-i(m)}\right)^{-1} X'_{-i(m)} Y_{-i}$

$X_{-i(m)}$ $X_{(m)}$ missing i-th row and Y_{-i} $\mathbf{Y}$ missing i-th observation.

Supposed residual JMA is

$$\hat{\varepsilon}_i(w) = \sum_{1}^{M} \widehat{w}_m \hat{\varepsilon}_{i(m)}$$

$$\text{or } \hat{\varepsilon}(w) = \sum_{1}^{M} \widehat{w}_m \bar{\varepsilon}_{(m)} = \bar{\boldsymbol{\varepsilon}} \boldsymbol{w} \; ; \; \bar{\boldsymbol{\varepsilon}} = (\bar{\varepsilon}_{(1)}, \dots, \bar{\varepsilon}_{(M)}) \text{ is matrix in } n \text{ x } M.$$

Residual sum squared of JMA is

$$CV(w) = \bar{\varepsilon}(w)' \bar{\varepsilon}(w) = w' \bar{\varepsilon} \bar{\varepsilon}' w$$

JMA estimator of μ is obtained by minimizing

$$\widehat{w} = \text{argmin } CV_n(w)$$

2 Mallows Model Averaging (MMA)

Mallows criterion is asymptotically proven similar to squared error criterion. Thus, the MMA estimator achieves the lowest possible squared error asymptotically.

Mallows criterion to *model averaging* is

$$C(\boldsymbol{W}) = \hat{\varepsilon}(\boldsymbol{W})'\hat{\varepsilon}(\boldsymbol{W}) + 2\sigma^2\boldsymbol{W}'\Phi$$

where $\hat{\varepsilon}(\boldsymbol{W}) = Y - X_M\widehat{\Theta}$ is the weighted residual for the weighted estimator and $\Phi = (\phi_1, \phi_2, ..., \phi_M)'$

It can be expressed as

$$C(\boldsymbol{W}) = \mathbf{W}'\hat{\varepsilon}\hat{\varepsilon}'\boldsymbol{W} + 2\sigma^2\boldsymbol{W}'\Phi$$

This is the quadratic form in vector *w*. So the weight $\widehat{\boldsymbol{W}}$ which minimizing Mallow criteria is the one that minimizing C(W) where w_m is element of $H_n^* = \left\{0, \frac{1}{N}, \frac{2}{N}, \dots, 1\right\} \subset H_n$

$$H_n = \left\{ w \epsilon [0,1]^M; \sum_1^M w_m = 1 \right\}$$

And *N* is integer.. Thus, *model averaging* estimator with Mallow criteria is

$$\widehat{W} = \underset{w_i \epsilon H_n^*}{\arg\min}\, C(\boldsymbol{W})$$

Construction Of Model Candidate

The performance of model averaging estimators could not be separated from candidate model construction . According to (Ando & Li, 2014), the candidate models can be provided in various way and sometimes depended on the field of study. In economics, business and finance, the candidate could be based on the formerly theory. Most of those study assume that candidate models are based on different competing theories for prediction model which sometimes are influenced by subject knowledge or expert theories.

Some of the construction independed on subjective perception are purposed by (Hansen B. E., 2007) and (Hansen & Racine, 2012). The candidate models is constructed by forming nested models of the data. Those models then are estimated to build the MA estimator.

Supposed set of candidate models $M = \{M_1,...,Mi\}$;

m-th model using first *m* element of x_i ; So the *m*th model

$$y_i = \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon$$

(Ando & Li, 2014) sugested a different perspective of the construction. The candidate is structured by the value of correlation between each covariate and response variable. The covariates with the same value form a design matrix of one model. It is clamed as an objective way in selecting the variables to construct a model. Supposed the marginal correlation between each predictor variable and the response variable is estimated by

$$\hat{\gamma} = n^{-1}X'y$$

Sorting the set of *p* regressors based on the marginal correlation magnitude to obtain *M* design matrix for *M* candidate model. The remaining variables not included are dropped.

Another approach of forming candidate model is purposed by (Magnus, Powell, & Prufer, 2010). All covariates are categorized in focus and auxiliary variables. The focus ones must be included in the model, called it main model. The candidate models are constructed by entering the combination of auxiliary variables to the main model.

Refer to equation (1), it can be written in

$$\boldsymbol{y} = X_1\beta_1 + X_2\beta_2 + \boldsymbol{\varepsilon}$$

Where $X = (X_1 \mid X_2)$ where X_1 is focus variables and X_2 is auxiliary variable

Each *i* candidate model can be written as

$$\boldsymbol{y_i} = X_1\beta_1 + C_i\beta_i$$

Where C_i is *i*-th combination of auxiliary civariates.

MODEL ESTIMATION

According to (Hansen & Racine, 2012), the estimators of the candidate models could include linear least squares, ridge regression, near neighbor estimators, series estimators and spline estimators. Most of the researchers, however, restrict to linear least squares estimation methods, such as (Hansen & Racine, 2012) and (Ando & Li, 2014) even in the existance of heteroscedasticity.

If β is the least squares estimator then for each model M_i in equation (2) on section 2

$$\hat{\beta}_i = (X_i'X_i)^{-1}X_i'y$$

Consequently,

$$\widehat{\boldsymbol{\mu}}_i = X_i\widehat{\boldsymbol{\beta}}_i = X_i(X_i'X_i)^{-1}X_i'y = P_i y$$

where

$$P_i = X_i(X_i'X_i)^{-1}X_i'$$

acts as projection matrix of the *i*-th model.

(Liu, R, & A, 2013) using generalized least squares (GLS) methods to estimate each of candidate model in economics application. Supposed matrix variance of the model in equation (1) in section 2 is

$$\Omega = diag(\sigma_1^2, \dots, \sigma_p^2)$$

The GLS estimator of each β_i can be written as

$$\hat{\beta}_{GLS} = (X_i'\Omega^{-1}X_i)^{-1}X_i'\Omega^{-1}y$$

Consequently, the GLS estimator for μ is

$$\hat{\boldsymbol{\mu}}_{GLS} = X_i(X_i'\Omega^{-1}X_i)^{-1}X_i'\Omega^{-1}y$$

CONCLUSION

There are some unsolved issues according to frequentist model averaging, to get a better performance in prediction and also to obtain easier computation process. Some of them had mention above, are follows: model weight criteria selection , the way of construction the candidate model and methods of model estimation.

REFERENCES

Ando, T., & Li, K.-C. (2014). A Model-Averaging Approach for Hig Dimensional Regression. *Journal of the American Statistical Association*, 254-265.

Burnham, K., & Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoritic Approach.* Berlin: Spinger-Verlag.

Claeskens, G., & Hjort, N. (2008). *Model Selection and Model Averaging*. NY: Cambridge University.

Drapper, N., & Smith, H. (1966). *Applied Regression Analysis.* Canada: John Wiley & Sons.

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized likelihood and Its Oracle Properties. *American Statistical Assosiation No. 96*, 1348-1360.

Gusnanto, A., & Pawitan, Y. (2014). Sparse Alternatives to Ridge Regression: A Random effects Approach. *Journal of Applied Statistics*.

Hansen, B. (2014). Model Averaging, Asumptotic Risk and Regressor Groups. *Quantitative Economics 5*, 495-530.

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica, 75*, 1175-1189.

Hansen, B., & Racine, J. (2012). Jackknife Model Averaging. *Jounral of Econometrics, 167*, 34-38.

Liu, Q., R, O., & A, Y. (2013). *Generalized Least Squares Model Averaging*. Kyoto: Kyoto University.

Magnus, J., Powell, & Prufer, P. (2010). A Comparison of Two Model Averaging techniques with Application to Growth Empirics. *Journal of Econometrics 154*, 139-153.

Raftery, A., Madigan, D., & Hoeting, J. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association Vol 92 No 437*, 179-191.

Tibhsirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B Vol 58 No.1*, 267-288.