Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 513-518
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

A Review on High-Dimensional Frequentist Model Averaging

Peipei Fu[1,2], Juming Pan[3*]

[1]School of Health Care Management, Shandong University, Jinan, China
[2]Key Laboratory of Health Economics and Policy Research, NHFPC (Shandong University), Jinan, China
[3]Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, USA
Email: *jpan@d.umn.edu

How to cite this paper: Fu, P.P. and Pan, J.M. (2018) A Review on High-Dimensional Frequentist Model Averaging. *Open Journal of Statistics*, **8**, 513-518.
https://doi.org/10.4236/ojs.2018.83033

Received: April 23, 2018
Accepted: June 10, 2018
Published: June 13, 2018

Abstract

Model averaging has attracted increasing attention in recent years for the analysis of high-dimensional data. By weighting several competing statistical models suitably, model averaging attempts to achieve stable and improved prediction. To obtain a better understanding of the available model averaging methods, their properties and the relationships between them, this paper is devoted to make a review on some recent progresses in high-dimensional model averaging from the frequentist perspective. Some future research topics are also discussed.

Keywords

Model Averaging, High-Dimensional Regression Models, Stable Prediction

1. Introduction

With the advent of high-throughput technologies, high-dimensional data have been frequently generated for the understanding of biological processes such as disease occurrence and cancer study. Motivated by these important applications, there has been a dramatic development in the statistical analysis of high-dimensional data; see [1] and [2], and examples therein.

Model selection and model averaging are two approaches used to improve estimation and prediction in the regression problems. Model selection assigns the weight of a single optimal model to 1 and weights for other candidate models to 0, thus the parsimonious and compact representations of the data can be obtained. In recent years, the shrinkage methods have become popular as they can achieve simultaneous model selection and parameter estimation. Such methods include, but are not limited to, the least absolute shrinkage and selection opera-

tor (LASSO, Tibshirani [3]), the smoothly clipped absolute deviation (SCAD, Fan and Li [4]), the elastic net (Zou and Hastie [5]), and the minimax concave penalty (MCP, Zhang [6]).

However, the process of model selection ignores the additional uncertainty or even introduces bias, and therefore often underestimates variance [7]. In addition, different selection methods or criteria may yield different best models. Hence inference based on the final model can be seriously misleading.

Instead of relying on only one model, model averaging compromises across a set of competing models by assigning different weights. In doing so, model uncertainty is incorporated into the conclusions about the unknown parameters. Besides, if the weights can be properly determined, then prediction performance could be enhanced [8].

Regarding model averaging techniques, Frequentist Model Averaging (FMA) and Bayesian Model Averaging (BMA) are two different methods in the literature. Compared with FMA, there are extensive references on BMA where a prior probability to each candidate model is set for the model uncertainty; for an overview of BMA, see [9]. On the other hand, the FMA approach, whose estimators are totally determined by data, is starting to receive more attention over the last decade, as the procedure avoids problems such as how to set priors and how to deal with the priors when they are in conflict.

The aim of this paper is to make a review on the current methods of the FMA in the high-dimensional linear models. The methods on FMA estimation are surveyed in Section 2. Some future research topics are discussed in Section 3.

2. High-Dimensional FMA

So far, most current model averaging approaches are developed for the classic setting in which the number of observations is greater than the number of predictors, with the main focus of determination of the weights for individual models. These approaches include Akaike information criterion model averaging (AIC, Akaike [10]), Bayesian information criterion model averaging (BIC, Hoeting *et al.* [11]), Mallows model averaging (Hansen [12]; Wan *et al.* [13]), and Jackknife model averaging (Hansen and Racine [14]; Zhang *et al.* [15]), to name but a few.

However, for the high-dimensional setting, model averaging has only recently been studied. This is very different from the finite dimensional case because many of the fixed dimensional model averaging procedures either do not work at all or, for their implementation, require some theoretical or computational adjustment.

Given the dataset of n observations, a linear regression model takes the form of

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

where y_i is the response in the ith trial, $x_{i1}, \cdots, x_{ip}$ are the predictors, $\beta_1, \cdots, \beta_p$ are the regression coefficients, and ϵ_i is the error term. Alternative-

ly, in matrix form, model (1) can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Developing for the data in which the number of predictors p is much greater than the number of observations n, Ando and Li [16] proposed a two-stage model averaging procedure. The procedure first divides p predictors into $M+1$ groups by the absolute marginal correlations between all predictors and the response. Let model M_k consist of the regressors with marginal correlations falling into the kth group. The first group has the highest values, and the $M+1$ group has values closest to 0 and is then discarded. Thus the number of candidate models is M. Each model can also be written in matrix form $\boldsymbol{y} = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$, for $k = 1, \cdots, M$. Given candidate models whose number of predictors is smaller than the sample size, the regression coefficients are estimated by the usual least-squares method as $\hat{\boldsymbol{\beta}}_k = \left(\boldsymbol{X}_k'\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k'\boldsymbol{y}$ and the predicted value $\hat{\boldsymbol{u}}_k = \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k$.

After the candidate models and their corresponding least-squares predicted values $\left\{\hat{\boldsymbol{u}}_1, \cdots, \hat{\boldsymbol{u}}_M\right\}$ are obtained, the second stage of procedure of [16] is to determine the model weights. Let $\tilde{\boldsymbol{u}}_k = \left(\tilde{\boldsymbol{u}}_k^{(-1)}, \cdots, \tilde{\boldsymbol{u}}_k^{(-n)}\right)$ be an n-dimensional vector, where $\tilde{\boldsymbol{u}}_k^{(-i)}, i = 1, \cdots, n$ is the predicted value of the ith observation from M_k using the data without the ith observation, then the optimal weight vector $\boldsymbol{w}$ is optimized by minimizing the delete-one cross-validation criterion

$$\hat{\boldsymbol{w}} = \mathrm{CV}\left(\boldsymbol{w}\right) = \left(\boldsymbol{y} - \tilde{\boldsymbol{u}}\right)'\left(\boldsymbol{y} - \tilde{\boldsymbol{u}}\right), \tag{3}$$

where $\tilde{\boldsymbol{u}} = \sum_{k=1}^{M} w_k\tilde{\boldsymbol{u}}_k$. Finally, the model averaging predicted value $\hat{\boldsymbol{u}}$ is expressed as

$$\hat{\boldsymbol{u}} = \sum_{k=1}^{M}\hat{w}_k\hat{\boldsymbol{u}}_k. \tag{4}$$

There are several contributions of Ando and Li [16]. One notable feature of this method is the relaxation on the total model weights. The standard constraint of the model weights summing up to 1 is relaxed to the model weights can be vary freely between 0 and 1, and it is shown that this relaxation is helpful to lower the prediction error. Furthermore, the algorithm is computationally feasible for high-dimensional data, since each candidate model and its corresponding weight are first determined in the low-dimensional setting and then organically combined. Theoretically, it is proved that the proposed method could asymptot-

ically achieve the lowest possible prediction loss, which is an important property in prediction performance.

Following [16], Ando and Li [17] further extended model averaging to high-dimensional generalized linear models. Still allowing the weights to alter between 0 and 1, the Kullback-Leibler distance is used in [17] as a replacement of the squared error for risk measure, to overcome several technical and theoretical challenges.

Nevertheless, Lin *et al.* [18] showed through a simulated example that the two-stage model averaging procedure in [16] tends to have high variance and may lead the final estimator to be overfitting. They argued that the increase in variance is due to the reuse of the same data for generating candidate models and estimating model weights in the two steps.

To reduce the variance of estimators, Lin *et al.* [18] proposed a random splitting approach by first dividing the original data set into training set D_{train}^{b} and test set D_{test}^{b} for B times, $b=1,\cdots,B$. For each D_{train}^{b}, the variable selection method LASSO is applied to determine candidate model $\hat{M}_{\lambda_k}$ for each candidate tuning parameter $\lambda_k, k=1,\cdots,K$ and the corresponding coefficients $\hat{\boldsymbol{\beta}}^{\left(b,\hat{M}_{\lambda_k}\right)}$. In the next step, the second level data $\left(y_i, z_{ik}\right)$ is constructed, $i=1,\cdots,n$, where

$$z_{ik}=\frac{\sum_{b\in I_i} x_i'\hat{\boldsymbol{\beta}}^{\left(b,\hat{M}_{\lambda_k}\right)}}{\left|I_i\right|}, \tag{5}$$

where I_i is the set of indexes of test dataset D_{test}^{b} that contain observation *i*. After z_{ik} is determined, the optimal weight vector $\boldsymbol{w}$ is estimated by minimizing

$$\sum_{i=1}^{n}\left(y_i-\sum_{k=1}^{K}w_k z_{ik}\right)^2. \tag{6}$$

Finally, the model averaging predicted value takes the form of

$$\hat{\boldsymbol{u}}=\frac{1}{B}\sum_{k=1}^{K}\left(\hat{w}_k\sum_{b=1}^{B}\hat{\boldsymbol{u}}^{\left(b,\hat{M}_{\lambda_k}\right)}\right). \tag{7}$$

The procedure of [18] selects candidate models and obtains estimators using training sets, while finds optimal weights using only test sets, which could successfully avoid model overfitting and could improve prediction accuracy by combining models from multiple random splits. The main price one pays for using the random splitting, however, is in significantly increased computational complexity.

3. Conclusion and Discussion

In this paper, we have made a review on the development of the FMA approach for high-dimensional linear regression models. The performance of the FMA procedures highly depends on how to choose weights in estimation, since different weights will result in different risks and asymptotic properties. Conse-

quently, much of the current work focuses on weight choice to achieve stable prediction. Another issue is how to deal with the high-dimensional settings as the least-squares estimates are not unique. The general idea is to reduce the dimensions first and then to combine the low-dimensional models using the appropriate weights.

Although substantial progress has been made recently, the research on the FMA approach is a relatively new topic, for which a lot of problems remain unsolved and future work still needs to be done.

On possible direction is the extension of the FMA approach to other modeling settings containing generalized linear mixed model and Cox proportional hazards model, both of which are widely used in biological and medical research. For example, Zhang and Zou [19] proposed a model averaging approach in linear mixed-effects models. To determine the optimal weight with more complex model structures still is a meaningful work.

We also note that missing values are quite common in high-dimensional data, which leaves space for further research in model averaging. Schomaker *et al.* [20] suggested an FMA method with the presence of missing observations for low-dimensional data. Imputation handling will also be considered in addressing missing data in the future study for larger data sets.

Finally, in current research on the weight choice, many focus on developing those weights which are non-negative; it seems interesting to explore the possibility of further relaxing the weights to allow for negative values. These and many other unsettled issues deserve further investigation.

Acknowledgements

The authors are grateful for a grant from Shandong University (IFYT18032).

References

[1] Bühlmann, P. and van de Geer, S. (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Berlin. https://doi.org/10.1007/978-3-642-20192-9

[2] Li, X. and Xu, R. (2009) High-Dimensional Data Analysis in Cancer Research. Springer, Berlin. https://doi.org/10.1007/978-0-387-69765-9

[3] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B*, **58**, 268-288.

[4] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[5] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*: *Series B*, **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[6] Zhang, C. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[7] Wang, H., Zhang, X. and Zou, G. (2009) Frequentist Model Averaging Estimation: A Review. *Journal of Systems Science and Complexity*, **22**, 732-748.

https://doi.org/10.1007/s11424-009-9198-y

[8] Liang, H., Zou, G., Wan, A.T.K. and Zhang, X. (2011) Optimal Weight Choice for Frequentist Model Average Estimators. *Journal of the American Statistical Association*, **106**, 1053-1066. https://doi.org/10.1198/jasa.2011.tm09478

[9] Fragoso, T.M. and Neto, F.L. (2015) Bayesian Model Averaging: A Systematic Review and Conceptual Classification. arXiv:1509.08864.

[10] Akaike, H. (1979) A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. *Biometrika*, **66**, 237-242. https://doi.org/10.1093/biomet/66.2.237

[11] Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1999) Bayesian Model Averaging. *Statistical Science*, **14**, 382-401.

[12] Hansen, B.E. (2007) Least Squares Model Averaging. *Econometrica*, **75**, 1175-1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x

[13] Wan, A., Zhang, X. and Zou, G. (2010) Least Squares Model Averaging by Mallows Criterion. *Journal of Econometrics*, **156**, 277-283. https://doi.org/10.1016/j.jeconom.2009.10.030

[14] Hansen, B.E. and Racine, J. (2012) Jackknife Model Averaging. *Journal of Econometrics*, **167**, 38-46. https://doi.org/10.1016/j.jeconom.2011.06.019

[15] Zhang, X.Y., Wan, A.T.K. and Zou, G.H. (2013) Model Averaging by Jackknife Criterion in Models with Dependent Data. *Journal of Econometrics*, **174**, 82-94. https://doi.org/10.1016/j.jeconom.2013.01.004

[16] Ando, T. and Li, K.C. (2014) A Model-Averaging Approach for High-Dimensional Regression. *Journal of the American Statistical Association*, **109**, 254-265. https://doi.org/10.1080/01621459.2013.838168

[17] Ando, T. and Li, K.C. (2017) A Weight-Relaxed Model Averaging Approach for High-Dimensional Generalized Linear Models. *The Annals of Statistics*, **45**, 2654-2679. https://doi.org/10.1214/17-AOS1538

[18] Lin, B., Wang, Q., Zhang, J. and Pang, Z. (2017) Stable Prediction in High-Dimensional Linear Models. *Statistics and Computing*, **27**, 1401-1412. https://doi.org/10.1007/s11222-016-9694-6

[19] Zhang, X., Zou, G. and Liang, H. (2014) Model Averaging and Weight Choice in Linear Mixed-Effects Models. *Biometrika*, **101**, 205-218. https://doi.org/10.1093/biomet/ast052

[20] Schomaker, M., Wan, A.T.K. and Heumann, C. (2010) Frequentist Model Averaging with Missing Observations. *Computational Statistics and Data Analysis*, **54**, 3336-3347. https://doi.org/10.1016/j.csda.2009.07.023