Jrl Syst Sci & Complexity (2009) 2009: 732–748

FREQUENTIST MODEL AVERAGING ESTIMATION: A REVIEW*

Haiying WANG · Xinyu ZHANG · Guohua ZOU

Received: 2 July 2009 / Revised: 2 September 2009

Abstract In applications, the traditional estimation procedure generally begins with model selection. Once a specific model is selected, subsequent estimation is conducted under the selected model without consideration of the uncertainty from the selection process. This often leads to the underreporting of variability and too optimistic confidence sets. Model averaging estimation is an alternative to this procedure, which incorporates model uncertainty into the estimation process. In recent years, there has been a rising interest in model averaging from the frequentist perspective, and some important progresses have been made. In this paper, the theory and methods on frequentist model averaging estimation are surveyed. Some future research topics are also discussed.

Key words Adaptive regression, asymptotic theory, frequentist model averaging, model selection, optimality.

1 Introduction

Traditional data analysis generally includes two stages: The first is to select an appropriate model, and the second is to make inference under the selected model as if this model had been given in advance. In reality, this process ignores the additional uncertainty or even bias introduced by model selection procedure and thus often underreports variance (or mean squared error) or reports too optimistic coverage probability of claimed confidence interval. An alternative to this process is model averaging. A model averaging estimator compromises across a set of competing models, and in doing so, incorporates model uncertainty into the conclusions about the unknown parameters. Besides, model averaging estimator often improves the risk in estimation because it provides a kind of insurance against selecting a very poor model.

The concept of model averaging, including frequentist model averaging (FMA) and Bayesian model averaging (BMA), appeared in about 1960s and most of the early papers focused on economic fields; see, for example, [1–5]. Compared with the FMA approach, there has been an enormous literature on the use of the BMA approach where the uncertainty on model is considered by setting a prior probability to each candidate model. Some examples include [6–10]. But, as commented by Hjort and Claeskens[11], in using the BMA approach, there exist some problems such as how to set prior probabilities and how to deal with the priors when they are in conflict with each other. In contrast, the FMA approach requires no priors and the

Haiying WANG · Xinyu ZHANG · Guohua ZOU
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, *China.*
Email: ghzou@amss.ac.cn.
*This research is supported by the National Natural Science Foundation of China under Grant Nos. 70625004, 10721101, and 70221001.

corresponding estimators are totally determined by data. Therefore, the FMA approach has received much attention over last decade; see, for example, [11–17].

The application of the FMA approach can be dated back to the work of Bates and Granger[1] on forecast combination. Recently, with the development of theory on this approach, it has been used in many fields. For instance, Claeskens, Croux and ven Kerckhoven[18] applied the approach to a study of diabetic retinopathy; Kapetanios, Labhard, and Price[19] used it to forecast UK inflation; Pesaran, Schleicher, and Zaffaroni[20] discussed its use as a way dealing with the risk of inadvertently using false models in portfolio management; and Wan and Zhang[21] applied it to tourism research. More extensively, the monograph by Claeskens and Hjort[22] presents many empirical illustrations on the use of the technique. Our current paper is devoted to make a review on the theory and specific operating methods of the FMA approach.

This paper is organized as follows. We first introduce the definition of the FMA estimators in Section 2, and then summarize the asymptotic theories of the FMA estimators in Section 3 and discuss the choice methods of the weights in the FMA estimators in Section 4. We consider model averaging based on various regression procedures in Section 5. Some future research topics are included in Section 6.

2 Definition of FMA Estimators

In this section, we take a linear model as an example to illustrate the definition of the FMA estimator. The extension to more general framework is straightforward. Consider the following linear model:

$$y = X\beta + Z\gamma + \varepsilon, \tag{1}$$

where $y(n \times 1)$ is a vector of response variable, $X(n \times p)$ and $Z(n \times q)$ are the non-random regressor matrices, $\beta(p \times 1)$ and $\gamma(q \times 1)$ are the parameter vectors, and $\varepsilon(n \times 1)$ is a random error vector. We assume that (X, Z) has full column rank $p+q$. Here, X contains the variables that must be included in the model fitting, while Z contains the "doubtful" variables that may be included in the model.

Clearly, by setting some components of $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_q)^t$ to be zeros, there are totally 2^q candidate models. Assume that our purpose is to estimate β. Let $\widehat{\beta}_{(S)}$ be the estimator of β under the candidate model S, where S is a subset of $\{1, 2, \cdots, q\}$, and the candidate model S means the model with β and γ_j $(j \in S)$ as unknown parameters. The traditional data analysis method takes the selected model as the one given in advance, and reports only the variance or mean squared error of $\widehat{\beta}_{(S)}$ if the candidate model S is selected, while the actual estimator is

$$\widehat{\beta} = \begin{cases} \widehat{\beta}_{(S_1)}, & \text{if the first model (say, } S_1\text{) is selected,} \\ \vdots & \vdots \\ \widehat{\beta}_{(S_{2^q})}, & \text{if the } 2^q\text{-th model (say, } S_{2^q}\text{) is selected.} \end{cases}$$

For simplicity, we rewrite the above estimator as

$$\widehat{\beta} = \sum_S \overline{\lambda}(S \mid \text{data})\widehat{\beta}_S,$$

where

$$\overline{\lambda}(S \mid \text{data}) = \begin{cases} 1, & \text{if the candidate model } S \text{ is selected;} \\ 0, & \text{otherwise,} \end{cases}$$

and 'data' means that the weights are determined based on the data and no priors are involved, which is also a difference between the FMA and BMA approaches.

The above estimator is the usual pre-test estimator which is not continuous and thus inadmissible (see [23]). So naturally, we consider the smoothed weights $\lambda(S|\text{data})$ and accordingly, the model averaging estimator is given by

$$\widehat{\beta} = \sum_S \lambda(S|\text{data})\widehat{\beta}_S,$$

where $0 \le \lambda(S|\text{data}) \le 1$, and $\sum_S \lambda(S|\text{data}) = 1$. Such an estimator is referenced as the FMA estimator of β which integrates the model selection and estimation precedure.

Now, we consider more general case. Assume that the quantity of interest is μ, and its estimator is $\widehat{\mu}_S$ under the candidate model S. Then, the FMA estimator of μ is given by

$$\widehat{\mu} = \sum_S \lambda(S \mid \text{data})\widehat{\mu}_S.$$

3 Asymptotic Theory of Frequentist Model Averaging Estimators

In this section, we first discuss asymptotic theory of the FMA estimators under parametric models, then consider its generalization to the case of semiparametric models.

3.1 Asymptotic Theory of the FMA Estimators under Parametric Models

Suppose data come from a model with the density

$$f_{\text{true}} = f(y, \beta, \gamma) = f(y, \beta_0, \gamma_0 + \delta/\sqrt{n}), \tag{2}$$

where $\beta(p \times 1)$ consists of parameters in all candidate models and β_0 is its true value, $\gamma(q \times 1)$ is a vector around γ_0 with perturbation $\delta/\sqrt{n}$, and γ_0 is a known vector, determined by the statistical problem of interest. This is a local misspecification framework with the null model when $\delta = 0$. This framework is crucial for deriving asymptotic results and some arguments on it can be found in [24–25]. By setting some components of δ to be zeros, there exist totally 2^q candidate models. Let the quantity of interest be $\mu_{\text{true}} = \mu(f_{\text{true}})$, where $\mu(\cdot)$ is a known function. Under the candidate model S (the model with β and δ_j, $j \in S$, as unknown parameters), the estimator of μ_{true} is given by

$$\widehat{\mu}_S = \mu(\widehat{\beta}_S, \widehat{\gamma}_S, \gamma_{0,\,S^C}),$$

where $\widehat{\beta}_S$ and $\widehat{\gamma}_S$ are the maximum likelihood estimators of β and γ_S based on the model, respectively, and S^C is the complement of S.

Hjort and Claeskens[11] studied the asymptotic properties of the FMA estimator with the following form

$$\widehat{\mu}_{H1} = \sum_S \lambda(S|D_n)\widehat{\mu}_S, \tag{3}$$

where $\lambda(S|D_n)$ is a weight function, and $D_n = \widehat{\delta}_{\text{full}}$ is an estimator of δ under the full model. Before arriving at the main results, let us introduce some notations. Denote the score function by

$$\begin{pmatrix} U(y) \\ V(y) \end{pmatrix} = \begin{pmatrix} \partial \log f(y, \beta_0, \gamma_0)/\partial\beta \\ \partial \log f(y, \beta_0, \gamma_0)/\partial\gamma \end{pmatrix}, \tag{4}$$

whose $(p+q)\times(p+q)$ variance matrix at the null point $(\beta_0^t, \gamma_0^t)^t$ is

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \text{ with inverse } J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}. \tag{5}$$

Let π_S be the projection matrix mapping $\delta = (\delta_1, \delta_2, \cdots, \delta_q)^t$ to the sub-vector $\pi_S\delta = \delta_S$ with the components δ_j $(j \in S)$. The following theorem involves asymptotic results of the FMA estimator $\widehat{\mu}_{H1}$.

Theorem 1 *If each weight $\lambda(S|\cdot)$ has at most a countable number of discontinuities, then*

$$\sqrt{n}(\widehat{\mu}_{H1} - \mu_{\text{true}}) \xrightarrow{d} \Lambda = \mu_\beta^t J_{00}^{-1}\zeta + \omega^t\{\delta - \widehat{\delta}(D)\}, \tag{6}$$

where '$\xrightarrow{d}$' denotes convergence in distribution, $D\sim N_q(\delta, \Psi)$ is the limit of D_n, $\mu_\beta = \frac{\partial\mu}{\partial\beta}|_{\beta=\beta_0,\gamma=\gamma_0}$, $\zeta \sim N_p(0, J_{00})$ is independent of D, $\widehat{\delta}(D) = \big\{\sum_S \lambda(S|D)\pi_S^t(\pi_S\Psi^{-1}\pi_S^t)^{-1}\pi_S\big\}\Psi^{-1}D$, $\Psi = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, and $\omega = J_{10}J_{00}^{-1}\mu_\beta - \frac{\partial\mu}{\partial\gamma}|_{\beta=\beta_0,\gamma=\gamma_0}$. The mean and variance of Λ are $\omega^t\{\delta - E\widehat{\delta}(D)\}$ and $\mu_\beta^t J_{00}^{-1}\mu_\beta + \omega^t\text{Var}\left(\widehat{\delta}(D)\right)\omega$, respectively.

Let $|S|$ be the number of the elements in S. Hjort and Claeskens[11] also studied the Akaike Information Criterion (AIC)[26] based on the above framework and gave the following result

$$\text{AIC}_{n,S} = D_n^t\,\Psi^{-1}(\pi_S\,\Psi^{-1}\pi_S^t)^{-1}\,\Psi^{-1}D_n - 2|S| + o_P(1), \tag{7}$$

which means that the $\text{AIC}_{n,S}$ is approximately determined by S and the estimator of δ under the full model. This indicates that the model selection estimator by AIC and the averaging estimators with Akaike weights in [12] and [15] are actually special cases of the estimator in (3). Further, the authors pointed out that the coverage probability of the confidence interval introduced by Burnham and Anderson[15] is biased, and then they constructed a confidence interval with confidence limits listed below and showed that such an interval has an asymptotic confidence level precisely the same as the intended level $1-2F_{\text{norm}}(z_u)$ ($F_{\text{norm}}(\cdot)$ is the standard normal distribution function):

$$\begin{aligned} \text{low}_n &= \widehat{\mu}_{H1} - \widehat{\omega}^t[D_n - \widehat{\delta}(D_n)D_n]/\sqrt{n} - z_u\widehat{\kappa}/\sqrt{n}, \\ \text{up}_n &= \widehat{\mu}_{H1} - \widehat{\omega}^t[D_n - \widehat{\delta}(D_n)D_n]/\sqrt{n} + z_u\widehat{\kappa}/\sqrt{n}, \end{aligned}$$

where $\widehat{\omega}$ and $\widehat{\kappa}$ are consistent estimators of ω and $\kappa = \sqrt{\mu_\beta^t J_{00}^{-1}\mu_\beta + \omega^t\,\Psi\omega}$, respectively, and z_u is the u-th standard normal quantile.

3.2 Asymptotic Theory of the FMA Estimators Under Semiparametric Models

Following the above asymptotic theory under parametric models, Hjort and Claeskens[25] considered the FMA approach under Cox's hazard regression models

$$h(u) = h_0(u)\exp(X\beta + Z\gamma) = h_0(u)\exp(X\beta + Z\delta/\sqrt{n}), \tag{8}$$

where, unlike the model (2), the null model corresponds to $\gamma = 0$ but not $\gamma = \gamma_0$ for simplicity. Note that some notations in current section share the same definitions as those before, except that they are defined based on the model (8). For a given parameter of interest $\mu(\beta, \gamma, H_0)$, where H_0 is the cumulative baseline hazard rate, the FMA estimator is defined as

$$\widehat{\mu}_{H2} = \sum_S \lambda(S|D_n)\widehat{\mu}_S, \tag{9}$$

where $\widehat{\mu}_S = \mu(\widehat{\beta}_S, \widehat{\gamma}_S, 0_{S^C}, \widehat{H}_{0,S})$, and $\widehat{H}_{0,S}$ is Aalen-Breslow type estimator of H_0 under the candidate model S. Let

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$$

be the limit of the second order derivatives of the partial likelihood with respect to β and γ (normalized by n^{-1} and calculated at the null point $(\beta_0^t, 0^t)^t$). Then the asymptotic results of the FMA estimator $\widehat{\mu}_{H2}$ can be provided as follows.

Theorem 2 *Suppose ordinary regularity conditions for Cox's model hold, and assume that for each S, the random weight $\lambda(S|D_n)$ used in (9) is such that $\lambda(S|D_n)$ tends in distribution to $\lambda(S|D)$, in terms of the limit D of D_n, where each $\lambda(S|\cdot)$ has at most a finite number of discontinuities. Then*

$$\sqrt{n}(\widehat{\mu}_{H2} - \mu_{\text{true}}) \xrightarrow{d} \Lambda = \Lambda_0 + (\omega - \kappa_1)^t\{\delta - \widehat{\delta}(D)\}, \tag{10}$$

where $\Lambda_0 \sim N(0, \Pi_0^2)$ is independent of D, $\kappa_1 = \kappa_1(t) = \{J_{10}J_{00}^{-1}F_0(t) - F_1(t)\}\frac{\partial \mu}{\partial H_0}$, and the definitions of $F_0(t)$, $F_1(t)$, and Π_0 can be found in [25] (*as they are complicated, we omit them for saving space*).

Based on Theorem 2, Hjort and Claeskens[25] also gave a useful representation of AIC for Cox's model which is similar to Equation (7).

Furthermore, Claeskens and Carroll[27] showed that all of the results in [11] hold in a commonly used semiparametric model, given that the Fisher information matrix for parametric models is replaced by the semiparametric information bound for semiparametric models, and maximum likelihood estimators for parametric models are replaced by semiparametric efficient profile estimators. The semiparametric model in this paper is partially linear model with normal errors

$$Y_i = X_i^t\alpha + g(T_i) + \varepsilon_i = X_i^t(\beta^t, \delta^t/\sqrt{n})^t + g(T_i) + \varepsilon_i. \tag{11}$$

For simplicity, the authors considered the case of two candidate models: The full model with $\alpha_{\text{full}} = (\beta^t, \gamma^t)^t$ and the null model with $\alpha_{\text{null}} = (\beta^t, 0^t)^t$. Then the FMA estimator of the parameter of interest $\mu_{\text{true}} = \mu(\alpha)$ is given by

$$\widehat{\mu}_C = \lambda(\widehat{\delta}_{\text{full}})\mu(\widehat{\alpha}_{\text{full}}) + \{1 - \lambda(\widehat{\delta}_{\text{full}})\}\mu(\widehat{\alpha}_{\text{null}}), \tag{12}$$

where $\widehat{\alpha}_{\text{full}}$ and $\widehat{\delta}_{\text{full}}$ are the estimators of α and δ under the full model, respectively, and $\widehat{\alpha}_{\text{null}}$ is the estimator of α under the null model of (11). Based on the same regularity conditions as those used in a local likelihood setting where one wishes to obtain strong uniform consistency of the local likelihood estimators, they gave the following asymptotic distribution about averaging estimator.

Theorem 3 *Under the local misspecification assumption,*

$$\sqrt{n}\{\widehat{\mu}_C - \mu(\alpha)\} \xrightarrow{d} \Lambda = \lambda(D)\Lambda_{\text{full}} + \{1 - \lambda(D)\}\Lambda_{\text{null}}, \tag{13}$$

where D, Λ_{full}, and Λ_{null} are, respectively, the limiting variables of $\widehat{\delta}_{\text{full}}$, $\sqrt{n}\{\mu(\widehat{\alpha}_{\text{full}}) - \mu_{\text{true}}\}$ and $\sqrt{n}\{\mu(\widehat{\alpha}_{\text{null}}) - \mu_{\text{true}}\}$.

Partition the semiparametric information bound $\mathcal{S}(\alpha)$ as well as its inverse in the following way,

$$\mathcal{S}(\alpha) = \begin{pmatrix} S_{\beta\beta}(\alpha) & S_{\beta\gamma}(\alpha) \\ S_{\gamma\beta}(\alpha) & S_{\gamma\gamma}(\alpha) \end{pmatrix}, \quad \mathcal{S}^{-1}(\alpha) = \begin{pmatrix} S^{\beta\beta}(\alpha) & S^{\beta\gamma}(\alpha) \\ S^{\gamma\beta}(\alpha) & S^{\gamma\gamma}(\alpha) \end{pmatrix},$$

then we get

$$E(\Lambda) = \omega^t(\delta - E\{\lambda(D)D\}),$$
$$\mathrm{Var}(\Lambda) = \mu_\beta^t(S_{\beta\beta})^{-1}\mu_\beta + \omega^t \mathrm{Var}\{\lambda(D)D\}\omega,$$

where $\mu_\beta = \frac{\partial \mu}{\partial \beta}|_{\beta=\beta_0,\gamma=0}$ and $\omega = S_{\gamma\beta}S_{\beta\beta}^{-1}\mu_\beta - \frac{\partial \mu}{\partial \gamma}|_{\beta=\beta_0,\gamma=0}$. Subsequently, a confidence interval with the limits

$$\widehat{\mu} - \widehat{\omega}^t[\{1 - \lambda(\widehat{\delta}_{\mathrm{full}})\}\widehat{\delta}_{\mathrm{full}}]/\sqrt{n} \pm z_u\widehat{\kappa}_2/\sqrt{n},$$

has the probability of coverage converging to an intended level, where z_u is the u-th standard normal quantile as before and $\widehat{\kappa}_2^2$ is a consistent estimator of the variance under the full model $\kappa_2^2 = \mu_\beta^t S_{\beta\beta}^{-1}\mu_\beta + \omega^t S^{\gamma\gamma}\omega$.

More recently, Wang[28] generalized Hjort and Claeskens's[11] results to linear models with measurement errors, and semiparametric varying coefficient partially linear models with and without measurement errors in the linear part.

In addition, for the model (1) with normal error assumption, Magnus and Durbin[13] found that if only the full and null models are considered, then the problem of estimating the coefficients of interest by using the FMA estimator is equivalent to that of finding an optimal estimator of the vector of coefficients of no interest given a single observation from a normal distribution. This is an interesting result as it implies that irrespective of the model's structure, the optimal solution for the $N(\theta, \sigma^2 I_{q\times q})$ problem (where $I_{q\times q}$ is an identity matrix) is also optimal for all regression models with the same value of q. Danilov and Magnus[29] generalized the equivalence property to the case of multiple candidate models. Further, Danilov and Magnus[30] considered forecasting problem. On the other hand, Zou et al.[31] extended Magnus and Durbin's findings to the large sample non-normal errors case under the local misspecification framework of Hjort and Claeskens[11].

4 Weight Choice for the FMA Estimators

An important issue with the FMA estimators is how to choose weights in estimation. Different weights will result in different risks and asymptotic properties, so in this section we introduce some work related to weight choice and show that some weight choice methods are asymptotically optimal, i.e., asymptotically achieving the lowest squared error among a family of estimators.

4.1 Weight Choice Based on the Information Criterion

Buckland, Burnham, and Augustin[12] began with the following model averaging estimator of a parameter μ (assumed to be common to all models),

$$\widehat{\mu}_B = \sum_{k=1}^{K} \lambda_k \widehat{\mu}_k, \tag{14}$$

where the index k denotes the k-th candidate model, $\widehat{\mu}_k$ is the estimator of μ on the basis of the k-th candidate model, λ_k is the weight associated with $\widehat{\mu}_k$, and $\sum_{k=1}^{K}\lambda_k = 1$.

Assuming that the weights λ_k $(k = 1, 2, \cdots, K)$ are known constants and the correlation coefficients between the estimators of parameter under different models equal to one, they gave the formula of the variance for the estimator $\widehat{\mu}_B$ as follows

$$\mathrm{Var}(\widehat{\mu}_B) = \left\{ \sum_{k=1}^{K} \lambda_k \sqrt{\mathrm{Var}(\widehat{\mu}_k|b_k) + b_k^2} \right\}^2, \tag{15}$$

where b_k is the misspecification bias which arises in estimating μ under the k-th candidate model. This variance can be estimated by substituting $\mathrm{Var}(\widehat{\mu}_k|b_k)$ and b_k with $\widehat{\mathrm{Var}}(\widehat{\mu}_k|b_k)$ and $\widehat{b}_k = \widehat{\mu}_k - \widehat{\mu}_B$.

However, in practice, the weights have to be estimated, and so Buckland, Burnham, and Augustin[12] resorted to information criteria of the form

$$\boldsymbol{I}_k = -2\log(L_k) + \ell, \tag{16}$$

where L_k is the maximized likelihood function under the k-th model and ℓ is a penalty function of the number of parameters and/or the number of observations. They then recommended to use the following weights

$$\lambda_k = \frac{\exp(-\boldsymbol{I}_k/2)}{\sum_{i=1}^{K}\exp(-\boldsymbol{I}_i/2)}, \quad k = 1, 2, \cdots, K. \tag{17}$$

If $\ell = 2p$, where p is the number of parameters, $\boldsymbol{I}_k$ would be AIC and accordingly the estimator is called smoothed AIC estimator with Akaike weight. Although they did not conduct theoretical study on such estimators, they presented three numerical examples to demonstrate the virtue of them. The weights of the form (17) have been widely used in literature. Examples include [21] and [32–34].

Burnham and Anderson[15] worked further following Buckland, Burnham, and Augustin's[12] idea and introduced a formal procedure for inference under multimodels which they termed as the 'multimodel inference'. They stated that even if the weights are Akaike weights, the variance of the estimator can still be estimated by $\big\{\sum_{k=1}^{K}\lambda_k\sqrt{\mathrm{Var}(\widehat{\mu}_k|b_k) + (\widehat{\mu}_k - \widehat{\mu}_B)^2}\big\}^2$. They also proposed the methods on establishing unconditional confidence intervals, estimating the relative importance of variables and constructing set for the K-L best model.

4.2 Weight Choice Based on Mallows' Criterion

Hansen[17] discussed the model averaging in least squares estimation and proposed a method that selects the weights by minimizing Mallows' criterion[35]. The model Hansen considered is a homoskedastic linear regression:

$$y_i = \mu_i + e_i, \quad i = 1, 2, \cdots, n, \tag{18}$$

$$\mu_i = \sum_{j=1}^{\infty}\theta_j x_{ij}, \tag{19}$$

$$E(e_i|x_i) = 0, \tag{20}$$

$$E(e_i^2|x_i) = \sigma^2, \tag{21}$$

with $x_i = (x_{i1}, x_{i2}, \cdots)$ and the assumptions that $E\mu_i^2 < \infty$ and (19) converges in mean square. Consider a sequence of approximating models $k = 1, 2, \cdots$, where the k-th model uses the first ϕ_k elements of x_i with $0 < \phi_1 < \phi_2 < \cdots$. (Note that in all Hansen's work on model averaging, he termed candidate models as approximating models, since, as he remarked in [36], models should be viewed as approximations of data generating process. Therefore, we also use the term 'approximating' when discuss his work.) Thus, the k-th approximating model is given by

$$y_i = \sum_{j=1}^{\phi_k}\theta_j x_{ij} + e_i, \quad i = 1, 2, \cdots, n. \tag{22}$$

The corresponding approximating error is $\sum_{j=\phi_k+1}^{\infty}\theta_j x_{ij}$. Using matrix notation, the model (22) can be rewritten as

$$Y = X_k\Theta_k + e,$$

where $Y = (y_1, y_2, \cdots, y_n)^t$, X_k is an $n \times \phi_k$ matrix with the ij-th element x_{ij}, $\Theta_k = (\theta_1, \theta_2, \cdots, \theta_{\phi_k})^t$, and $e = (e_1, e_2, \cdots, e_n)^t$. Denote $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^t$. Let $K = K(n) \le n$ be the approximating model with the largest number of regressors and assume $X_K^t X_K$ to be invertible. Let $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_K)^t$ be a weight vector in the unit simplex in $\boldsymbol{R}^K$:

$$\boldsymbol{H}_n = \left\{\lambda \in [0,1]^K : \sum_{k=1}^{K}\lambda_k = 1\right\}, \tag{23}$$

then the least squares model averaging estimator of Θ_K is defined as

$$\widehat{\Theta}(\lambda) = \sum_{k=1}^{K}\lambda_k\begin{pmatrix}\widehat{\Theta}_k\\ 0\end{pmatrix}, \tag{24}$$

where $\widehat{\Theta}_k$ is the least squares estimator of Θ_k under the k-th approximating model. Correspondingly, the model averaging estimator of μ is

$$\widehat{\mu}(\lambda) = X_K\widehat{\Theta}(\lambda) = P(\lambda)Y, \tag{25}$$

where $P(\lambda) = \sum_{k=1}^{K}\lambda_k P_k$ and $P_k = X_k(X_k^tX_k)^{-1}X_k^t$ is the projection matrix under the k-th approximating model.

The Mallows' criterion for the model averaging estimator is

$$C_n(\lambda) = \big(Y - X_K\widehat{\Theta}(\lambda)\big)^t\big(Y - X_K\widehat{\Theta}(\lambda)\big) + 2\sigma^2\lambda^t\Phi, \tag{26}$$

where $\Phi = (\phi_1, \phi_2, \cdots, \phi_K)^t$. In applications, Hansen[17] proposed to choose weights by minimizing this criterion, i.e., the selected weight vector is

$$\widehat{\lambda} = \underset{\lambda\in\boldsymbol{H}n}{\operatorname{argmin}}\, C_n(\lambda). \tag{27}$$

Define the average squared error and expected conditional squared error as $L_n(\lambda) = (\widehat{\mu}(\lambda) - \mu)^t(\widehat{\mu}(\lambda) - \mu)$ and $R_n(\lambda) = E(L_n(\lambda)|x_1, x_2, \cdots, x_n)$, respectively. Let $\boldsymbol{H}_n(N)$ be the subset of $\boldsymbol{H}_n$, where λ_k's are restricted to the set $\{0, \frac{1}{N}, \frac{2}{N}, \cdots, 1\}$ with some integer N. Denote

$$\widehat{\lambda}(N) = \underset{\lambda\in\boldsymbol{H}_n(N)}{\operatorname{argmin}}\, C_n(\lambda). \tag{28}$$

The following theorem shows the asymptotic optimality of $\widehat{\mu}(\widehat{\lambda}(N))$ when the weights are constrained to $\boldsymbol{H}_n(N)$.

Theorem 4 *As $n \to \infty$, if*

$$\xi_n = \inf_{\lambda\in\boldsymbol{H}_n} R_n(\lambda) \to \infty \tag{29}$$

almost surely, and for some fixed integer $N < \infty$,

$$E(|e_i|^{4(N+1)}|x_i) \le c_e < \infty, \tag{30}$$

then

$$\frac{L_n(\widehat{\lambda}(N))}{\inf\limits_{\lambda\in\boldsymbol{H}_n(N)} L_n(\lambda)} \xrightarrow{p} 1, \tag{31}$$

where '$\xrightarrow{p}$' denotes convergence in probability, and c_e is a positive constant.

The above theorem shows that $\widehat{\mu}(\widehat{\lambda}(N))$ asymptotically achieves the lowest possible squared error when we constrain the weight vector to the discrete set $\boldsymbol{H}_n(N)$, i.e., is asymptotically optimal. It should be pointed out that Kabaila[37] demonstrated that post-model-selection estimators with asymptotically efficient property can have rather inefficient small sample performance, this finding is likely to carry over to the FMA estimators.

Furthermore, Hansen[38] proposed to use Mallows' model averaging method to do forecast and showed that the Mallows' criterion is an asymptotically unbiased estimator of both the in-sample mean squared error and the out-of-sample one-step-ahead mean squared forecast error. Hansen[39] studied least squares estimation of an autoregressive model with a root close to unity. He proposed two measures to evaluate the efficiency of the estimators: the asymptotic mean squared error and forecast expected squared error. Numerical comparison of Mallows' model averaging method with many other methods shows that Mallows' model averaging estimator often has smaller risk.

More recently, noting that all approximating models considered in [17] are nested and the asymptotic optimality shown in Theorem 4 is based on a discrete weight set, Wan, Zhang and Zou[40] improved results of [17] by removing these two restrictions (the removing of the second restriction is an open problem in [17]). They considered a non-nested framework by allowing regressors in the approximating model (22) to be any ϕ_k regressors belonging to x_i. Under some reasonable conditions, they obtained the asymptotic optimality of $\widehat{\mu}(\widehat{\lambda})$ as follows. Note that $\widehat{\lambda}$ defined in (27) is more general than $\widehat{\lambda}(N)$ in (28). Also, the condition (33) in the following theorem is stronger than the condition (29) in Theorem 4.

Theorem 5 *As $n \to \infty$, if for some integer $1 \le G < \infty$,*

$$E(|e_i|^{4G}|x_i) \le c_e^* < \infty, \tag{32}$$

and

$$K\xi_n^{-2G} \sum_{k=1}^{K} \Big(R_n(\lambda_k^o)\Big)^G \to 0, \tag{33}$$

then

$$\frac{L_n(\widehat{\lambda})}{\inf\limits_{\lambda\in\boldsymbol{H}_n} L_n(\lambda)} \xrightarrow{p} 1, \tag{34}$$

where c_e^* is a positive constant, and λ_k^o is a $K \times 1$ vector in which the k-th element is one and the others are zeros.

4.3 Weight Choice Based on the Cross-Validation Criterion

Hansen and Racine[41] proposed to select the weights of least squares model averaging estimator by minimizing a deleted-1 cross-validation criterion, and the method is termed as the jackknife model averaging (JMA). Compared with Mallows' model averaging method, this method is appropriate for more general linear models, i.e., the random errors are allowed to be with heteroskedastic variances $(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$. In addition, like [40], the approximating models are allowed to be non-nested. The other set-up of the model considered in [41] is the same as that in [17].

The jackknife version of the model averaging estimator of μ is

$$\widetilde{\mu}(\lambda) = \sum_{k=1}^{K} \lambda_k \widetilde{P}_k Y \triangleq \widetilde{P}(\lambda) Y,$$

where $\widetilde{P}_k = \widetilde{D}_k(P_k - I_n) + I_n$, $\widetilde{D}_k$ is the $n \times n$ diagonal matrix with the i-th diagonal element being $(1 - h_{ii}^k)^{-1}$, $h_{ii}^k = X_{k,i}(X_k^t X_k)^{-1} X_{k,i}^t$, and $X_{k,i}$ is the i-th row of X_k. Let the expected squared error of $\widetilde{\mu}(\lambda)$ be $\widetilde{R}_n(\lambda) = E\left(\mu - \widetilde{\mu}(\lambda)\right)^t \left(\mu - \widetilde{\mu}(\lambda)\right)$. The deleted-1 cross-validation criterion is defined as

$$CV(\lambda) = \big(Y - \widetilde{\mu}(\lambda)\big)^t \big(Y - \widetilde{\mu}(\lambda)\big), \tag{35}$$

and the JMA estimator is $\widehat{\mu}(\widehat{\lambda}^*)$ with the weights

$$\widehat{\lambda}^* = \underset{\lambda \in \boldsymbol{H}_n}{\operatorname{argmin}} \ CV(\lambda).$$

The following theorem builds the asymptotic optimality of the JMA estimator.

Theorem 6 *As $n \to \infty$, if the condition* (29) *and the following conditions hold almost surely:*

$$\inf_i \sigma_i^2 \geq c_\sigma > 0 \ \ \textit{with } c_\sigma \textit{ being a constant}; \tag{36}$$

$$\textit{for some integer } G, \ \sup_i E\big(e_i^{4G} | x_i\big) < \infty; \tag{37}$$

$$\textit{for two constants } 0 < c_1 < \infty \textit{ and } c_2 > 0, \ \phi_k \geq c_1 k^{c_2}; \tag{38}$$

$$\max_{1 \leq k \leq K} \max_{1 \leq i \leq n} h_{ii}^k \to 0, \tag{39}$$

then

$$\frac{L_n(\widehat{\lambda}^*)}{\inf\limits_{\lambda \in \boldsymbol{H}_n(N)} L_n(\lambda)} \xrightarrow{p} 1, \tag{40}$$

and

$$\frac{R_n(\widehat{\lambda}^*)}{\inf\limits_{\lambda \in \boldsymbol{H}_n(N)} R_n(\lambda)} \xrightarrow{p} 1. \tag{41}$$

Note that, compared with the asymptotic optimality in Theorems 4 and 5, Theorem 6 additionally contains the asymptotic optimality in the sense of achieving the lowest risk value.

4.4 Weight Choice Based on the Unbiased Estimator of Risk

Under the parametric models discussed in Section 3.1, by deriving the unbiased estimator of risk, Liang, Zou, and Zhang[42] proposed to choose weights by minimizing the following criterion:

$$WC = \{\widehat{\omega}^t \widehat{\Psi}^{1/2} (\widehat{a}^*(\boldsymbol{Z}_n) - \boldsymbol{Z}_n)\}^2 + 2\widehat{\omega}^t \widehat{\Psi}^{1/2} \frac{\partial \widehat{a}^*(\boldsymbol{Z}_n)}{\partial \boldsymbol{Z}_n^t} \widehat{\Psi}^{1/2} \widehat{\omega}, \tag{42}$$

where $\boldsymbol{Z}_n = \widehat{\Psi}^{-1/2} D_n$, $\widehat{a}^*(\boldsymbol{Z}_n) = \{\sum_S \lambda(S|\boldsymbol{Z}_n) \widehat{\Psi}_S\} \boldsymbol{Z}_n$, $\widehat{\Psi}_S = \widehat{\Psi}^{-1/2} \pi_S^t (\pi_S \widehat{\Psi}^{-1} \pi_S^t)^{-1} \pi_S \widehat{\Psi}^{-1/2}$, $\widehat{\omega}$ and $\widehat{\Psi}$ are the estimators of ω and Ψ, respectively, and all other notations are the same as

those in Section 3.1. The calculation of the partial derivative in (42) can be done by explicit formulas or numerical differential methods.

In particular, for non-random weights, i.e., $\widehat{a}^*(\boldsymbol{Z}_n) = \big\{\sum_S \lambda(S)\,\widehat{\Psi}_S\big\}\boldsymbol{Z}_n$, (42) reduces to

$$WC = \left\{\widehat{\omega}^t\,\widehat{\Psi}^{1/2}\left(\sum_S \lambda(S)\,\widehat{\Psi}_S - I_q\right)\boldsymbol{Z}_n\right\}^2 + 2\widehat{\omega}^t\,\widehat{\Psi}^{1/2}\left\{\sum_S \lambda(S)\,\widehat{\Psi}_S\right\}\widehat{\Psi}^{1/2}\widehat{\omega}.$$

For the simplest case where only the full and null models are taken into account, we have

$$WC = (\lambda_{\text{full}} - 1)^2\left(\widehat{\omega}^t\,\widehat{\Psi}^{1/2}\boldsymbol{Z}_n\right)^2 + 2\lambda_{\text{full}}\widehat{\omega}^t\,\widehat{\Psi}\widehat{\omega}.$$

Thus, the desired weights are given by

$$\lambda_{\text{full}} = 1 - \frac{\widehat{\omega}^t\,\widehat{\Psi}\widehat{\omega}}{(\widehat{\omega}^t\,\widehat{\Psi}^{1/2}\boldsymbol{Z}_n)^2} \quad \text{and} \quad \lambda_{\text{null}} = \frac{\widehat{\omega}^t\,\widehat{\Psi}\widehat{\omega}}{(\widehat{\omega}^t\,\widehat{\Psi}^{1/2}\boldsymbol{Z}_n)^2}. \tag{43}$$

It can be seen that the weights obtained in (43) are not exactly the same as those in (5.5) of Hjort and Claeskens[11]. The reason is that Hjort and Claeskens[11] minimized the risk function itself, whereas Liang, Zou, and Zhang[42] minimized its unbiased estimator. However, they are close. In fact, when only δ is the unknown quantity, by noting that $E(\omega^t\,\Psi^{1/2}\boldsymbol{Z})^2 = \omega^t\,\Psi\omega + (\omega^t\,\Psi^{-1/2}\delta)^2$ with $\boldsymbol{Z} = \Psi^{-1/2}D$, the numerator and denominator of weights in Hjort and Claeskens[11] are just the expectations of those in (43), respectively. On the other hand, it can be shown that $\sqrt{n}(\widehat{\mu}_{\text{full}} - \widehat{\mu}_{\text{null}}) \approx -\widehat{\omega}^t\,\widehat{\Psi}^{1/2}\boldsymbol{Z}_n$. Thus, the weights in (43) are just the James-Stein type weights studied in [43].

On the basis of the criterion given in (42), we can assess the performance of a weight form by calculating its WC value, and compare different weight forms accordingly.

The criterion given in (42) is based on the large sample theory of Hjort and Claeskens[11]. For the small sample case, assuming the model framework in Section 2, Liang et al.[44] derived a criterion that can also be used to compare different weight forms.

5 Model Averaging Based on Various Regression Procedures

In the previous sections, the structures of candidate models for averaging across are generally the same. In fact, for the case with different structures of candidate models, model averaging is also possible. In 2001, Yang[14] gave an algorithm named 'Adaptive Regression by Mixing (ARM)' which can result in a weighted estimator that combines the estimators from different regression procedures. Here the regression procedures can be dramatically different, for example, including simple linear regression, additive modeling, projection pursuit, and neural network, etc. Consider the following regression setting

$$Y_i = f(X_i) + \sigma(X_i)\cdot\varepsilon_i,$$

where $(X_i, Y_i)_{i=1}^n$ are i.i.d. from the joint distribution of (X, Y) with $Y = f(X) + \sigma(X)\cdot\varepsilon$. The explanatory variable X could be multidimensional with the distribution P_X, and the error term ε is assumed to be independent of X and has the density $\varsigma(t)$ with mean 0 and a finite variance. The goal is to estimate the regression function f based on the data $Z^n = (X_i, Y_i)_{i=1}^n$. Let ϱ_k, $1 \le k \le K$, denote the proposed regression procedures. Also, let $\widehat{f}_{k,i}(x; Z^i)$ and $\widehat{\sigma}_{k,i}(x) = \widehat{\sigma}_{k,i}(x; Z^i)\,(i \ge 1)$ denote the estimators of f and σ_k by the procedure ϱ_k based on

Z^i, respectively. Then the algorithm of the ARM can be described below (n is assumed to be even for simplicity):

Step 0 Randomly permute the order of the observations.

Step 1 Split the data into two parts $Z^{(1)} = (X_i, Y_i)_{i=1}^{n/2}$ and $Z^{(2)} = (X_i, Y_i)_{i=n/2+1}^{n}$.

Step 2 Obtain the estimates $\widehat{f}_{k,n/2}(x; Z^{(1)})$ of f based on $Z^{(1)}$ for $1 \le k \le K$. Estimate the variance function $\sigma^2(x)$ by $\widehat{\sigma}^2_{k,n/2}(x)$.

Step 3 For each k, evaluate predictions. For $n/2 + 1 \le k \le n$, predict Y_i by $\widehat{f}_{k,n/2}(X_i)$. Compute

$$E_k = \frac{\prod_{i=n/2+1}^{n} \varsigma\big((Y_i - \widehat{f}_{k,n/2}(X_i))/\widehat{\sigma}_{k,n/2}(X_i)\big)}{\prod_{i=n/2+1}^{n} \widehat{\sigma}_{k,n/2}(X_i)}.$$

Step 4 Compute the current weight for procedure ϱ_k. Let $\lambda_k = \frac{E_k}{\sum_{k=1}^{K} E_k}$.

Step 5 Repeat Steps 0–4 $(U-1)$ more times and average the weights over the U random permutations. Let $\widehat{\lambda}_k$ denote the obtained weight of procedure ϱ_k. The final estimator is

$$\widehat{f}_n(x) = \sum_{k=1}^{K} \widehat{\lambda}_k \widehat{f}_{k,n}(x). \tag{44}$$

Many experiments have been done in the paper to show the performance of such an estimator. For analyzing the risk property of this kind of estimators, they constructed a similar but more general model averaging estimator as follows.

For each n, choose an integer N_n of order n that satisfies $1 \le N_n \le n$. Let $\lambda_{k,n-N_n+1} = \varpi_k, k = 1, 2, \cdots$, where $\varpi_k' s$ are positive numbers that sum to 1. For $n - N_n + 2 \le i \le n$, let

$$\lambda_{k,i} = \frac{\varpi_k \prod_{l=n-N_n+1}^{i-1} \varsigma\Big(\big(Y_{l+1} - \widehat{f}_{k,l}(X_{l+1})\big)/\big(\widehat{\sigma}_{k,l}(X_{l+1})\big)\Big)/\widehat{\sigma}_{k,l}(X_{l+1})}{\sum_{k=1}^{\infty} \varpi_k \prod_{l=n-N_n+1}^{i-1} \varsigma\Big(\big(Y_{l+1} - \widehat{f}_{k,l}(X_{l+1})\big)/\big(\widehat{\sigma}_{k,l}(X_{l+1})\big)\Big)/\widehat{\sigma}_{k,l}(X_{l+1})}. \tag{45}$$

Note that $\sum_{k\ge 1} \lambda_{k,i} = 1$ for each $i = n - N_n + 1, \cdots, n$. Let

$$\widetilde{f}_i(x) = \sum_k \lambda_{k,i} \widehat{f}_{k,i}(x) \tag{46}$$

and define

$$\overline{\widehat{f}}_n(x) = \frac{1}{N_n} \sum_{i=n-N_n+1}^{n} \widetilde{f}_i(x), \tag{47}$$

and use ϱ^* to denote the procedure producing $\{\overline{\widehat{f}}_n, n \ge 1\}$.

Define the risk of a procedure ϱ for estimating f at the sample size n as $R(f; n; \varrho) = E||f - \widehat{f}_n||^2$ with the expectation taken under the regression function f and the variance function σ^2, where $\|f - g\| = (\int |f(x) - g(x)|^2 dP_X)^{1/2}$. Consider the following two conditions:

A1 The regression function $f(x)$ is uniformly bounded ($||f||_\infty \le A < \infty$), and $\sigma(x)$ is uniformly bounded above and below ($0 < \underline{\sigma} \le \sigma(x) \le \overline{\sigma} < \infty$). Estimators produced by ϱ_k also satisfy these two requirements;

A2 The error density ς is such that for each pair $0 < s_0 < 1$ and $T > 0$, there exists a constant B (depending on s_0 and T) such that

$$\int h(x) \log \frac{\varsigma(x)}{(1/s)\varsigma((x-t)/s)} \mu(dx) \le B\big((1-s)^2 + t^2\big)$$

for all $s_0 \le s \le s_0^{-1}$ and $-T < t < T$.

Theorem 7 below shows the bound of risk of the model averaging estimator, which gives an important justification for the use of the ARM.

Theorem 7 *If ς is known with mean* 0 *and variance* 1*, and the conditions* A1 *and* A2 *hold, then for any given countable collection of estimation procedures $\Delta = \{\varrho_k, k \ge 1\}$, the estimator $\widehat{\widetilde{f}}_n(x)$ resulted from the procedure ϱ^* has the following properties*

$$R(f; n; \varrho^*) \le C_1 \inf_k \Big(\frac{1}{N_n} \log \frac{1}{\varpi_k} + \frac{C_2}{N_n} \sum_{l=n-N_n+1}^{n} \big(E||\sigma^2 - \widehat{\sigma}^2_{k,l}||^2 + E||f - \widehat{f}_{k,l}||^2 \big) \Big), \tag{48}$$

where the constant C_1 depends on A and $\widehat{\sigma}$, and C_2 depends on A, $\frac{\overline{\sigma}}{\sigma}$, and ς. This upper bound also applies to the average risk of $\widetilde{f}_i$, $n - N_n + 1 \le i \le n$, that is,

$$\frac{1}{N_n} \sum_{i=n-N_n+1}^{n} E||f - \widetilde{f}_i||^2 \le C_1 \inf_k \Big(\frac{1}{N_n} \log \frac{1}{\varpi_k} + \frac{C_2}{N_n} \sum_{l=n-N_n+1}^{n} (E||\sigma^2 - \widehat{\sigma}^2_{k,l}||^2 + E||f - \widehat{f}_{k,l}||^2) \Big).$$

Risk bound in estimating variance function and the corresponding algorithm are also provided in Section 6 of [14], we omit them for saving space.

With homoscedastic errors assumption, Yang[45] weakened the two conditions A1 and A2, and more importantly, gave explicit constants in the risk bound. The model considered is

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{id})$, $f(\cdot)$ is the true regression function, and ε_i is the random error with mean 0 and variance σ^2. Models to be selected or averaged are those with the subsets of the explanatory variables from $\{X_1, X_2, \cdots, X_d\}$. The risk of an estimator is defined as $R(f, \widehat{f}) = E||f - \widehat{f}||^2$. The general form of the method then can be presented as follows.

For $i = n/2+1$, let $\lambda_{k,i} = 1/K$ (the number of regression procedures) and for $n/2+1 < i \le n$, let

$$\lambda_{k,i} = \frac{(\widehat{\sigma}_{k,n/2})^{-(i-n/2-1)} \exp\Big(-\frac{1}{2\widehat{\sigma}^2_{k,n/2}} \sum_{l=n/2+1}^{i-1} \big(Y_l - \widehat{f}_{k,n/2}(\boldsymbol{X}_l)\big)^2 \Big)}{\sum_{j=1}^{K} (\widehat{\sigma}_{j,n/2})^{-(i-n/2-1)} \exp\Big(-\frac{1}{2\widehat{\sigma}^2_{j,n/2}} \sum_{l=n/2+1}^{i-1} \big(Y_l - \widehat{f}_{j,n/2}(\boldsymbol{X}_l)\big)^2 \Big)}. \tag{49}$$

Then define

$$\widetilde{\lambda}_k = \frac{2}{n} \sum_{i=n/2+1}^{n} \lambda_{k,i} \tag{50}$$

and let

$$\widetilde{f}_n(x) = \sum_{k=1}^{K} \widetilde{\lambda}_k \widehat{f}_{k,n/2}(x) \tag{51}$$

be the estimator.

The two weakened conditions are:

A3 There exists a constant $\tau > 0$ such that for all $i \ge 1$, with probability one, we have

$$\sup_{k \ge 1} ||\widehat{f}_{k,i} - f||_\infty \le \sqrt{\tau}\sigma.$$

A4 There exist constants $0 < \xi_1 \le 1 \le \xi_2 < \infty$ such that

$$\xi_1 \le \frac{\widehat{\sigma}^2_{k,i}}{\sigma^2} \le \xi_2$$

with probability one for all $k \ge 1$ and $i \ge 1$.

For simplicity, they gave only the result with Gaussian errors.

Theorem 8 *Assume that the errors are Gaussian and that the conditions* A3 *and* A4 *are satisfied. Then the risk of the combined regression estimator $\widetilde{f}_n$ satisfies*

$$E||\widetilde{f}_n - f||^2 \le (1 + \xi_2 + 9\tau/2)\inf_{k\ge 1}\Big(\frac{4\sigma^2 \log K}{n} + \frac{1}{\xi_1}E||\widehat{f}_{k,n/2} - f||^2 + \frac{C(\xi_1,\xi_2)}{\sigma^2}E(\widehat{\sigma}^2_{k,n/2} - \sigma^2)^2\Big),$$

where $C(\xi_1,\xi_2) = \frac{1/\xi_1 - 1 + \log \xi_2}{\xi_1^2(1/\xi_2 - 1)^2}$.

Yuan and Yang[46] went further along Yang's[45] idea and proposed a method named the adaptive regression by mixing with model screening (ARMS), which combines models selected by some criteria instead of all the sub-models in which some bad models may be included. Here we present a general version of the ARMS method (a version that is easy in computation can be found in [46]) and give the risk bound of the estimator derived from this method. Let Γ_s be a reduced list of candidate models based on any consideration using half of the data. Then the weights and estimators are similar to (49)–(51) except that "$k \in \{1, 2, \cdots, K\}$" is replaced by "$k \in \Gamma_s$": Let $k \in \Gamma_s$ with the size K_s. For $i = n/2 + 1$, define $\lambda_{k,i} = 1/K$ and for $n/2 + 1 \le i \le n$, define

$$\lambda_{k,i} = \frac{(\widehat{\sigma}_{k,n/2})^{-(i-n/2-1)} \exp\Big(-\frac{1}{2\widehat{\sigma}^2_{k,n/2}}\sum_{l=n/2+1}^{i-1}\big(Y_l - \widehat{f}_{k,n/2}(\boldsymbol{X}_l)\big)^2\Big)}{\sum_{j\in\Gamma_s}(\widehat{\sigma}_{j,n/2})^{-(i-n/2-1)} \exp\Big(-\frac{1}{2\widehat{\sigma}^2_{j,n/2}}\sum_{l=n/2+1}^{i-1}\big(Y_l - \widehat{f}_{j,n/2}(\boldsymbol{X}_l)\big)^2\Big)}. \tag{52}$$

Then define

$$\widetilde{\lambda}_k = \frac{2}{n}\sum_{i=n/2+1}^{n}\lambda_{k,i} \tag{53}$$

and let

$$\widetilde{f}(x) = \sum_{k\in\Gamma_s}\widetilde{\lambda}_k\widehat{f}_{k,n/2}(x) \tag{54}$$

be the estimator.

Theorem 9 *Assume that the errors are Gaussian and the conditions* A3 *and* A4 *are satisfied. Then for any $k \in \Gamma$, the risk of the combined regression estimator $\widetilde{f}$ using the ARMS satisfies*

$$\begin{aligned}E||\widetilde{f} - f||^2 \le& \tau\sigma^2 P(k \notin \Gamma_s) + \\ &(1 + \xi_2 + 9\tau/2)\Big(\frac{2\sigma^2 E\log K_s}{n} + \frac{1}{\xi_1}E||\widehat{f}_k - f||^2 + \frac{C(\xi_1,\xi_2)}{\sigma^2}E(\widehat{\sigma}^2_{k,n/2} - \sigma^2)^2\Big).\end{aligned}$$

In particular, when K_s is upper bounded by a constant K_0, we have

$$\begin{aligned}E||\widetilde{f} - f||^2 \le& \tau\sigma^2 P(k_* \notin \Gamma_s) + \\ &(1 + \xi_2 + 9\tau/2)\Big(\frac{2\sigma^2 E\log K_0}{n} + \frac{1}{\xi_1}E||\widehat{f}_{k_*} - f||^2 + \frac{C(\xi_1,\xi_2)}{\sigma^2}E(\widehat{\sigma}^2_{k_*,n/2} - \sigma^2)^2\Big),\end{aligned}$$

where k_* is the model index in Γ that the corresponding model minimizes the risk.

6 Future Researches

In this paper, we have made a review on the development of the FMA approach. Asymptotic theory of the FMA estimators has been presented and the choice methods of the weights in the FMA estimators have been summarized. Although great progress has been made recently, the research on the FMA approach is a relatively new topic, for which a lot of issues remain unsolved and further development is needed.

One meaningful work is the generalization of the FMA approach to more complex models such as the generalized varying coefficient semiparametric partially linear models and transformation models in survival analysis so that the approach can be applied widely in practice. Accordingly, the optimal choice methods of weight should be investigated. In fact, this issue has not been completely addressed even for the parametric models. For example, as we mentioned in Section 4.4, Liang, Zou, and Zhang[42] suggested a criterion of choosing weights. But the proposed criterion requires a large sample size. What is the corresponding criterion when the sample size is small? The problem remains unsolved.

In current research on the weight choice, many focus on developing those weights which have some optimality properties, but no corresponding methods have been provided to estimate the risks of the resultant estimators. Also, note that in order to evaluate the performance of the FMA approach, the squared error loss is often utilized. No doubt, other loss functions including asymmetric loss functions like the LINEX loss is worthy of investigation. On the other hand, what will happen if different estimation methods are used? For instance, in Section 3.1, what Hjort and Claeskens[11] considered is the maximum likelihood estimation. How about the M-estimator?

A problem with the FMA approach arises when the dimension of the optional or doubtful part of the parameters in the model is high. For example, if there are 10 optional parameters, then, generally, $2^{10} = 1024$ candidate models should be considered, and thus the FMA approach will usually be very time consuming. Recently, using the equivalence theorem introduced by Magnus and Durbin[13], Magnus, Powell and Prüfer[9] developed a WALS (weighted-average least squares) estimator based on Laplace prior. After a transformation, the required computing time for the WALS estimator is linear with respect to the dimension of the optional parameters rather than exponential. However, the method in [9] is based on Bayesian point of view. How to release from the computation burden for the FMA approach is a challenge research topic.

Another interesting question is whether the FMA approach is applicable to complex data. [25] and [28] have shown that it can be used in modeling censored data and the data with measurement errors. Schomaker, Wan and Heumann[47] studied the FMA approach with missing observations and two explicit operational approaches were presented and compared. Building similar theory and methods for some other types of the missing data or other types of complex data like longitudinal data warrants future researches.

An important reason for adopting model averaging estimation is that the traditional data analysis method ignores the uncertainty produced by model selection. There is another approach that automatically incorporates uncertainty from selection stage by selecting variables and estimating parameters simultaneously. This approach is based on the penalized function such as SCAD[48] penalized regression and adaptive LASSO[49] methods and sometimes has "Oracle Properties[48]". The advantages and disadvantages of the FMA approach compared to this approach are still unknown and should be explored.

References

[1] J. M. Bates and C. M. J. Granger, The combination of forecasts, *Operations Research Quarterly*, 1969, **20**: 451–468.

[2] D. A. Bessler and J. A. Brandt, Forecasting livestock prices with individual and composite methods, *Applied Economics*, 1981, **13**: 513–522.

[3] R. T. Clemen and R. L. Winkler, Combining economic forecasts, *Journal of Business and Economic Statistics*, 1986, **4**: 39–46.

[4] P. Newbold and C. W. J. Granger, Experience with forecasting univariate time series and the combination of forecasts, *Journal of the Royal Statistical Society, Series A*, 1974, **2**: 131–165.

[5] R. F. Phillips, Composite forecasting: An integrated approach and optimality reconsidered, *Journal of Business & Economic Statistics*, 1987, **5**: 389–395.

[6] M. A. Clyde and E. George, Model uncertainty, *Statistical Science*, 2004, **19**: 81–94.

[7] D. Draper, Assessment and propagation of model uncertainty, *Journal of the Royal Statistical Society: Series B*, 1995, **57**: 45–70.

[8] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, Bayesian model averaging: A tutorial, *Statistical Science*, 1999, **14**: 382–417.

[9] J. R. Magnus, O. Powell, and P. Prüfer, A comparison of two averaging techniques with an application to growth empirics, *Journal of Econometrics*, 2009, in press, doi:10.1016/j.jeconom.2009.07.004.

[10] A. E. Reftery, D. Madigen, and J. A. Hoeting, Bayesian model averaging for regression models, *Journal of the American Statistical Association*, 1997, **92**: 179–191.

[11] N. L. Hjort and G. Claeskens, Frequestist model average estimators, *Journal of the American Statistical Association*, 2003, **98**: 879–899.

[12] S. T. Buckland, K. P. Burnham, and N. H. Augustin, Model selection: An integral part of inference, *Biometrics*, 1997, **53**: 603–618.

[13] J. R. Magnus and J. Durbin, Estimation of regression coefficients of interest when other regression coefficients are of no interest, *Econometrica*, 1999, **67**: 639–643.

[14] Y. Yang, Adaptive regression by mixing, *Journal of the American Statistical Association*, 2001, **96**: 574–586.

[15] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Infromation-Theoretic Approach*, Springer, New York, 2002.

[16] G. Leung and A. R. Barron, Infromation theory and mixing least-squares regressions, *Information Theory, IEEE Transactions*, 2006, **52**: 3396–3410.

[17] B. E. Hansen, Least squares model averaging, *Econometrica*, 2007, **75**: 1175–1189.

[18] G. Claeskens, C. Croux, and J. ven Kerckhoven, Variable selection for logit regression using a prediction-focused information criterion, *Biometrics*, 2006, **62**: 972–979.

[19] G. Kapetanios, V. Labhard, and S. Price, Forecasting using Bayesian and information-theoretic model averaging, *Journal of Business and Economic Statistics*, 2008, **26**: 33–41.

[20] M. H. Pesaran, C. Schleicher, and P. Zaffaroni, Model averaging in risk management with an application to futures markets, *Journal of Empirical Finance*, 2009, **16**: 280–305.

[21] A. T. K. Wan and X. Zhang, On the use of model averaging in tourism research, *Annals of Tourism Research*, 2009, **36**: 525–532.

[22] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, Cambridge University Press, New York, 2008.

[23] G. G. Judge and M. E. Bock, *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*, North-Holland, Amsterdam, 1978.

[24] N. L. Hjort and G. Claeskens, Rejoinder, *Journal of the American Statistical Association*, 2003, **98**: 938–945.

[25] N. L. Hjort and G. Claeskens, Focused information criteria and model averaging for the Cox hazard regression model, *Journal of the American Statistical Association*, 2006, **110**: 1449–1464.

[26] H. Akaike, Maximum likelihood identification of Gaussian autoregression moving average models, *Biometrika*, 1973, **60**: 255–265.

[27] G. Claeskens and R. J. Carroll, An asymptotic theory for model selection inference in general semiparametric problems, *Biometrika*, 2007, **94**: 249-265.
[28] H. Wang, Frequentist model averaging estimation, Master Thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 2009.
[29] D. Danilov and J. R. Magnus, On the harm that ignoring pretesting can cause, *Journal of Econometrics*, 2004, **122**: 27–46.
[30] D. Danilov and J. R. Magnus, Forecast accuracy after pretesting with an application to the stock market, *Journal of Forecasting*, 2004, **23**: 251–274.
[31] G. H. Zou, A. T. K. Wan, X. Wu, and T. Chen, Estimation of regression coefficients of interest when other regression coefficient are of no interest: The case of non-normal errors, *Statistics & Probability Letters*, 2007, **77**: 803–810.
[32] K. P. Burnham and D. R. Anderson, Multimodel inferenc uderstanding AIC and BIC in model selection, *Sociological Methods & Research*, 2004, **33**: 261–304.
[33] F. E. Turheimer, R. Hinz, and V. J. Cunningham, On the undecidability among kinetic models: From model selection to model averaging, *Journal of Cerbral Blood Flow & Metabolism*, 2003, **23**: 490–498.
[34] E. J. Wagenmakes and S. Farrell, AIC model selecion using Akaike weights, *Psychonomic Bulletin & Review*, 2004, **11**: 192–196.
[35] C. L. Mallows, Some comments on C_p, *Technometrics*, 1973, **15**: 661–675.
[36] B. E. Hansen, Challenges for econometric model selection, *Econometric Theory*, 2005, **21**: 60–68.
[37] P. Kabaila, On variable selection in linear regression, *Econometric Theory*, 2002, **18**: 913–925.
[38] B. E. Hansen, Least squares forecast averaging, *Journal of Econometrics*, 2008, **146**: 342–350.
[39] B. E. Hansen, Averaging estimators for autoregressions with a near unit root, *Journal of Econometrics*, 2009, forthcoming.
[40] A. T. K. Wan, X. Zhang, and G. Zou, Least squares model combining by Mallows criterion, Technical Report, Department of Management Sciences, City University of Hong Kong, 2009.
[41] B. E. Hansen and J. S. Racine, Jacknife model averaging, Technical Report, Department of Economics, University of Wisconsin-Madison, 2009.
[42] H. Liang, G. Zou, and X. Zhang, Choice of weights for frequentist model average estimators, Technical Report, Department of Biostatistics and Computational Biology, University of Rochester, 2009.
[43] T. H. Kim and H. White, James-Stein type estimators in large samples with application to the least absolute deviations estimator, *Journal of the American Statistical Association*, 2001, **96**: 697–705.
[44] H. Liang, G. Zou, A. T. K. Wan, and X. Zhang, On optimal weight choice in a frequentist model average estimator, Technical Report, Department of Biostatistics and Computational Biology, University of Rochester, 2009.
[45] Y. Yang, Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica*, 2003, **13**: 783–809.
[46] Z. Yuan and Y. Yang, Combining linear regression models: When and how? *Journal of the American Statistical Association*, 2005, **100**: 1202–1204.
[47] M. Schomaker, A. T. K. Wan, and C. Heumann, Frequentist model averaging with missing observations, *Computational Statistics and Data Analysis*, 2009, in press, doi:10.1016/j.csda.2009.07.023.
[48] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 2001, **96**: 1348–1360.
[49] H. Zou, The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association*, 2006, **101**: 1418–1429.