SCIENCE CHINA
Mathematics

• ARTICLES •

January 2012 Vol. 55 No. 1: 1–XX
doi: 10.1007/s11425-000-0000-0

Choice of weights in FMA estimators under general parametric models

Xinyu ZHANG[1,2,*], Guohua ZOU[1] & Hua LIANG[3]

[1] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190*, China;*
[2]*Center for Forecasting Science, Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York* 14642*, U.S.A.*

Email:xinyu@amss.ac.cn, ghzou@amss.ac.cn, hliang@bst.rochester.edu

Received November 17, 2011; accepted January 22, 2012; published online January 22, 2012

Abstract The choice of weights in frequentist model average estimators is an important but difficult problem. [15] suggested a criterion for the choice of weight under a general parametric framework which is termed as the generalized OPT (GOPT) criterion in the present paper. However, no properties and applications of the criterion have been studied. This paper is devoted to the further investigation of the GOPT criterion. We show how to use this criterion for comparison of some existing weights such as the smoothed AIC-based and BIC-based weights and for the choice between model averaging and model selection. Its connection to the Mallows and ordinary OPT criteria is built. The asymptotic optimality on the criterion in the case of non-random weights is also obtained. Finite sample performance of the GOPT criterion is assessed by simulations. Application to the analysis of two real data sets is presented as well.

Keywords asymptotic optimality, likelihood inference, model averaging, model selection, model selection diagnostics

MSC(2010) 62F10, 62F99, 62H12

Citation: Zhang X Y, Zou G H, Liang H. Choice of weights in FMA estimators under general parametric models. Sci China Math, 2012, 55, doi: 10.1007/s11425-000-0000-0

1 Introduction

Model averaging is a standard approach to deal with model uncertainty. Unlike model selection, model averaging combines across a set of competing models rather than relies on a single "final model". One important motivation behind model averaging is that somehow it avoids ignoring useful information from the form of the relationship between response and covariates [1]. Another is that it provides a kind of insurance against selecting a very poor model [14]. There is a large collection of literature on Bayesian model averaging. See, for example, [12] for a review. Yet studies from a frequentist perspective have been relatively few. However, some significant frequentist model averaging (FMA) strategies are developed in recent years. Examples include information criterion-based weighting [2, 3, 24], adaptive regression mixing [22, 23], and Mallows model averaging (MMA) [7, 19]. Besides, [9] first introduced a local misspecification framework for the study of the asymptotic properties of FMA estimators. See [20] for a review on FMA.

*Corresponding author

More recently, under linear regression models, [15] introduced a general random weight form which includs smoothed AIC (S-AIC), smoothed BIC (S-BIC) and many other weights as special cases. Based on this weight form, an optimal mean squares error average estimator (termed as OPT estimator) was developed. Furthermore, under general parametric models, using the results of [9], they introduced a criterion for selecting weights in the FMA estimators which is referenced as the generalized OPT (GOPT) criterion in the present article. However, they did not discuss the properties and applications of the GOPT criterion, except for a connection with the weights of [9]. The purpose of this paper is to further investigate the GOPT criterion. We will show that under linear regression with normal disturbances, the criterion is equivalent to the Mallows criterion of [7] and OPT criterion of [15] when the non-random and general random weight forms are used, respectively, by which we conclude that the GOPT criterion indeed extends the ordinary OPT criterion of weight choice and is a unified version of Mallows and OPT criteria. Moreover, focusing on the criterion with non-random weights, we will present an asymptotic optimality in the sense that the asymptotic risk of the resulting model average estimator is asymptotically equivalent to the infeasible optimal risk. We will also demonstrate via simulation studies and practical applications that in most of cases, a best weight can be chosen by minimizing GOPT criterion. In addition, based on the GOPT criterion, we will discuss a model selection diagnostics method.

The remainder of this paper is organized as follows. In Section 2, we introduce the model framework and the GOPT criterion developed by [15], and give the explicit forms of the criterion when S-AIC, S-BIC and smoothed focused information criterion (S-FIC) weights are used. In Section 3, we show the connection of this criterion to the Mallows and OPT criteria. In Section 4, an asymptotic optimality on the criterion in the case of non-random weights is presented. In Section 5, we conduct simulation experiments to assess the performance of the GOPT criterion. The criterion is also applied to the analysis of two real data sets in Section 6. In Section 7, we further discuss a model selection diagnostics method. We make some concluding remarks in Section 8. The appendix includes some technical details.

2 Model framework and criterion function

Assume that the observations $Y_1, ..., Y_n$ are i.i.d. and come from the density of the form

$$f_{\text{true}}(y) = f(y, \theta, \gamma) = f(y, \theta, \gamma_0 + \delta/\sqrt{n}), \tag{2.1}$$

where θ is a $p \times 1$ unknown vector, γ_0 is a $q \times 1$ known vector, and δ is a $q \times 1$ unknown vector representing the degree of the departure from the null model (the model contains only θ as unknown parameter vector). This local misspecification framework indicates that squared model bias is of size $O(1/n)$, the most possible large sample approximations. Some arguments related to this framework appear in [10] and [5]. The model (2.1) contains the full θ which is considered as the necessary parameter in the model based on theory or other grounds, and potentially partial elements of δ for which we are uncertain whether we should include in the model. The parameter of interest (or focused parameter) is $\mu = \mu(\theta, \gamma) = \mu(\theta, \gamma_0 + \delta/\sqrt{n})$. Clearly, there are 2^q sub-model estimators, each corresponding to the subset $S \subset \{1, ..., q\}$ in the sense of $\delta_j = 0$ for $j \in S^{\text{C}}$, where S^{C} is the complement of S. When the sub-model S is chosen, let $\hat{\theta}_S$ and $\hat{\gamma}_S$ denote the maximum likelihood estimators (MLEs) of the corresponding parameters, i.e., θ and the vector consisting of γ_j with $j \in S$. Therefore, the MLE of μ is $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^{\text{C}}})$.

Denote by J_{full} the $(p+q) \times (p+q)$ information matrix of the full model evaluated at the null point (θ, γ_0): $J_{\text{full}} = \text{var}_0 \begin{pmatrix} U(Y) \\ V(Y) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$ with the inverse $J_{\text{full}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}$, where $U(y) = \partial \log f(y, \theta, \gamma_0)/\partial\theta$ and $V(y) = \partial \log f(y, \theta, \gamma_0)/\partial\gamma$ are the score functions. As in [9], for regression models with Y_i coming from density $f_{i,\text{true}}(y_i \mid x_i) = f(y_i \mid x_i, \theta, \gamma) = f(y_i \mid x_i, \theta, \gamma_0 + \delta/\sqrt{n})$, a closely related

matrix

$$J_{n,\text{full}} = \frac{1}{n}\sum_{i=1}^{n} \text{var}_0 \begin{pmatrix} \partial \log f(Y_i \mid x_i, \theta, \gamma_0)/\partial\theta \\ \partial \log f(Y_i \mid x_i, \theta, \gamma_0)/\partial\gamma \end{pmatrix} = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix}$$

is assumed to converge to a suitable positive definite matrix J_{full} when n tends to infinity. Let π_S be the projection matrix mapping the vector $v = (v_1, ..., v_q)^t$ to its subvector $\pi_S v = v_S$ consisting of v_j with $j \in S$. Denote $K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, $K_S = (\pi_S K^{-1} \pi_S^t)^{-1}$, $H_S = K^{-1/2}\pi_S^t K_S \pi_S K^{-1/2}$, and $\omega = J_{10}J_{00}^{-1}\partial\mu/\partial\theta - \partial\mu/\partial\gamma$ with the partial derivatives evaluated at the null point (θ, γ_0). We define H_ϕ as the null matrix of size $q \times q$ (ϕ is the empty set).

Let $D_n = \hat{\delta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0)$. The FMA estimator of μ is written as $\hat{\mu} = \sum_S \bar{c}(S \mid D_n)\hat{\mu}_S$, where $\bar{c}(S \mid D_n)$'s are weights based on D_n. From [9], we see that $D_n \xrightarrow{d} D \sim N_q(\delta, K)$,

$$\sqrt{n}(\hat{\mu}_S - \mu) \xrightarrow{d} \Lambda_S \equiv \left(\frac{\partial\mu}{\partial\theta}\right)^t J_{00}^{-1}M + \omega^t(\delta - K^{1/2}H_S K^{-1/2}D), \tag{2.2}$$

and

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \Lambda \equiv \left(\frac{\partial\mu}{\partial\theta}\right)^t J_{00}^{-1}M + \omega^t\{\delta - \hat{\delta}(D)\}, \tag{2.3}$$

where $M \sim N_p(0, J_{00})$ is independent of D, $\hat{\delta}(D) = K^{1/2}\left\{\sum_S \bar{c}(S \mid D)H_S\right\}K^{-1/2}D \equiv K^{1/2}H\left(\bar{c}(D)\right)K^{-1/2}D$, and the vector $\bar{c}(\cdot)$ consists of $\bar{c}(S|\cdot)$'s. Thus, the asymptotic risk of $\hat{\mu}$ is given by

$$R_a(\hat{\mu}) = E(\Lambda^2) = \tau_0^2 + E\left(\omega^t\hat{\delta}(D) - \omega^t\delta\right)^2 \tag{2.4}$$

with $\tau_0^2 = \left(\frac{\partial\mu}{\partial\theta}\right)^t J_{00}^{-1}\left(\frac{\partial\mu}{\partial\theta}\right)$.

Denote $Z = K^{-1/2}D$. Then Z follows $N(a^*, I_q)$ with $a^* = K^{-1/2}\delta$. So we can write $\hat{\delta}(D) = K^{1/2}\left\{\sum_S \bar{c}(S \mid K^{1/2}Z)H_S\right\}Z \equiv K^{1/2}\left\{\sum_S c(S \mid Z)H_S\right\}Z = K^{1/2}H\left(c(Z)\right)Z \equiv K^{1/2}a^*(Z)$. [15] derived an unbiased estimator of $R_a(\hat{\mu})$,

$$\widehat{R}_a(\hat{\mu}) \quad = \tau_0^2 - \omega^t K\omega + \left\{\omega^t K^{1/2}(a^*(Z) - Z)\right\}^2 + 2\omega^t K^{1/2}\frac{\partial a^*(Z)}{\partial Z^t}K^{1/2}\omega. \tag{2.5}$$

Assuming that K and ω are known and removing the first two terms that are unrelated to the weights $c(S|Z)$, the right hand of (2.5) can be viewed as a focused criterion for the choice of weights in estimating the focused parameter μ, written as

$$\mathcal{W}^*(\hat{\mu}) \quad = \quad \left\{\omega^t K^{1/2}(a^*(Z) - Z)\right\}^2 + 2\omega^t K^{1/2}\frac{\partial a^*(Z)}{\partial Z^t}K^{1/2}\omega. \tag{2.6}$$

In particular, for non-random weights, i.e., $a^*(Z) = H(c)Z$, we obtain $\partial a^*(Z)/\partial Z^t = H(c)$. So (2.6) reduces to

$$\mathcal{W}_0(\hat{\mu}) = \left\{\omega^t K^{1/2}(H(c)Z - Z)\right\}^2 + 2\omega^t K^{1/2}H(c)K^{1/2}\omega. \tag{2.7}$$

The $\mathcal{W}^*(\hat{\mu})$ depends on some unknown quantities: ω, K, and Z, which can be replaced by their sample versions $\hat{\omega}$, $\hat{K}$, and $Z_n = \hat{K}^{-1/2}D_n$ on the basis of the full model, respectively, then the GOPT criterion is to minimize the following

$$\mathcal{W}(\hat{\mu}) = \left\{\hat{\omega}^t\hat{K}^{1/2}(\hat{a}^*(Z_n) - Z_n)\right\}^2 + 2\hat{\omega}^t\hat{K}^{1/2}\frac{\partial\hat{a}^*(Z_n)}{\partial Z_n^t}\hat{K}^{1/2}\hat{\omega}, \tag{2.8}$$

where a hat appears in $\hat{a}^*(Z_n)$ because $\hat{K}$ substitutes for K in $a^*(Z_n)$. The calculation of the partial derivative $\frac{\partial\hat{a}^*(Z_n)}{\partial Z_n^t}$ can be conducted by explicit formulas (for example, in the cases of the following $c_{\text{AIC}}(S \mid Z_n)$, $c_{\text{BIC}}(S \mid Z_n)$ and $c_{\text{FIC}}(S \mid Z_n)$ weights) or numerical differentiation.

On the basis of the criterion given in (2.8), we can compare different weights. Once again, our idea is to select the weights minimizing $\mathcal{W}(\hat{\mu})$. Denote by $\mid S \mid$ the number of the elements in S. If we use the S-AIC weights, which are defined as (c.f. [9])

$$c_{\text{AIC}}(S \mid Z_n) = \frac{\exp\{\frac{1}{2}AIC_{n,S}\}}{\sum_{S'}\exp\{\frac{1}{2}AIC_{n,S'}\}} \approx \frac{\exp\{\frac{1}{2}\left(Z_n^t\hat{H}_S Z_n - 2 \mid S \mid\right)\}}{\sum_{S'}\exp\{\frac{1}{2}\left(Z_n^t\hat{H}_{S'} Z_n - 2 \mid S' \mid\right)\}}, \tag{2.9}$$

then (2.8) reduces to

$$\begin{aligned}\mathcal{W}_{\text{S-AIC}} &= \left\{\hat{\omega}^t\hat{K}^{1/2}(\hat{H}(c_{\text{AIC}}(Z_n)) - I_q)Z_n\right\}^2 + 2\hat{\omega}^t\hat{K}^{1/2}\left\{\hat{H}(c_{\text{AIC}}(Z_n))\right.\\ &\quad \left.+\sum_S c_{\text{AIC}}(S \mid Z_n)\hat{H}_S Z_n Z_n^t\hat{H}_S - \hat{H}(c_{\text{AIC}}(Z_n))Z_n Z_n^t\hat{H}(c_{\text{AIC}}(Z_n))\right\}\hat{K}^{1/2}\hat{\omega},\end{aligned}$$

where a hat appears in $\hat{H}(c_{\text{AIC}}(Z_n))$ because $\hat{K}$ substitutes for K in $H(c_{\text{AIC}}(Z_n))$. If we use the S-BIC weights, which are defined as

$$c_{\text{BIC}}(S \mid Z_n) = \frac{\exp\{\frac{1}{2}BIC_{n,S}\}}{\sum_{S'}\exp\{\frac{1}{2}BIC_{n,S'}\}} \approx \frac{\exp\{\frac{1}{2}\left(Z_n^t\hat{H}_S Z_n - \mid S \mid \log n\right)\}}{\sum_{S'}\exp\{\frac{1}{2}\left(Z_n^t\hat{H}_{S'} Z_n - \mid S' \mid \log n\right)\}}, \tag{2.10}$$

then (2.8) reduces to $\mathcal{W}_{\text{S-BIC}}$, which has the same expression as $\mathcal{W}_{\text{S-AIC}}$ except that c_{AIC} is replaced by c_{BIC}. If we use the S-FIC weights, which are defined as

$$c_{\text{FIC}}(S \mid Z_n) = \frac{\exp\{-\frac{1}{2}\kappa FIC_{n,S}/\hat{\omega}^t\hat{K}\hat{\omega}\}}{\sum_{S'}\exp\{-\frac{1}{2}\kappa FIC_{n,S'}/\hat{\omega}^t\hat{K}\hat{\omega}\}}, \tag{2.11}$$

where $FIC_{n,S} = \{\hat{\omega}^t\hat{K}^{1/2}(I_q - \hat{H}_S)Z_n\}^2 + 2\hat{\omega}^t\hat{K}^{1/2}\hat{H}_S\hat{K}^{1/2}\hat{\omega}$ and $\kappa \geqslant 0$ is an algorithmic parameter, then (2.8) reduces to

$$\begin{aligned}\mathcal{W}_{\text{S-FIC}} &= \left\{\hat{\omega}^t\hat{K}^{1/2}(\hat{H}(c_{\text{FIC}}(Z_n)) - I_q)Z_n\right\}^2 + 2\hat{\omega}^t\hat{K}^{1/2}\left\{\hat{H}(c_{\text{FIC}}(Z_n))\right.\\ &\quad \left.-\kappa\sum_S c_{\text{FIC}}(S \mid Z_n)\hat{H}_S Z_n A_S + \kappa\hat{H}(c_{\text{FIC}}(Z_n))Z_n\sum_S c_{\text{FIC}}(S \mid Z_n)A_S\right\}\hat{K}^{1/2}\hat{\omega},\end{aligned}$$

where $A_S = (\hat{\omega}^t\hat{K}\hat{\omega})^{-1}\hat{\omega}^t\hat{K}^{1/2}(I_q - \hat{H}_S)Z_n\hat{\omega}^t\hat{K}^{1/2}(I_q - \hat{H}_S)$.

3 Connection to Mallows and OPT criteria

For the linear model setting, the Mallows and OPT criteria of weight choice are suggested by [7] and [15], respectively. We will show in this section that under the normality condition, the result (2.8) generalizes and unifies these two criteria. More specifically, we consider the following linear model

$$Y_i = x_i^t\beta + u_i^t\gamma + \varepsilon_i, \quad i = 1, ..., n, \tag{3.1}$$

where β is a $\tilde{p}\times 1$ vector, $\gamma = \delta/\sqrt{n}$ is a $\tilde{q}\times 1$ vector, and $\varepsilon_i \sim N(0, \sigma^2)$. It is easily verified that

$$J_{n,\text{full}} = \frac{1}{\sigma^2}\begin{pmatrix}\Sigma_{00} & 0 & \Sigma_{01}\\ 0 & 2 & 0\\ \Sigma_{10} & 0 & \Sigma_{11}\end{pmatrix},$$

where $\Sigma_{00} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^t$, $\Sigma_{01} = \frac{1}{n}\sum_{i=1}^{n} x_i u_i^t$, and $\Sigma_{11} = \frac{1}{n}\sum_{i=1}^{n} u_i u_i^t$.Write $Y = (Y_1, ..., Y_n)^t$, $X = (x_1, ..., x_n)^t$ and $U = (u_1, ..., u_n)^t$. Assume that $(X : U)$ has full column rank and let

$$\hat{\sigma}^2 = Y^t \left\{ I_n - (X : U) \left((X : U)^t (X : U) \right)^{-1} (X : U)^t \right\} Y/(n - \tilde{p} - \tilde{q}).$$

Then we have

$$\hat{K} = \hat{\sigma}^2(\Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01})^{-1} = n\hat{\sigma}^2(U^t M U)^{-1}, \tag{3.2}$$

where $M = I_n - X(X^t X)^{-1} X^t$.

We now simply assume $c(S|Z_n)$ to be non-random and rewrite $\hat{H}(c(Z_n))$ as $\hat{H}(c)$. Let $h_i = x_i^t\beta + u_i^t\gamma$ be the parameter of interest. Then $\omega = U^t X(X^t X)^{-1} x_i - u_i \equiv v_i$ and thus the criterion function (2.8) can be written as

$$\begin{aligned} \mathcal{W}(\hat{h}_i) &= \left\{ \left(\hat{H}(c) Z_n - Z_n \right)^t \hat{K}^{1/2} v_i \right\}^2 + 2 v_i^t \hat{K}^{1/2} \hat{H}(c) \hat{K}^{1/2} v_i \\ &= \left(\hat{H}(c) Z_n - Z_n \right)^t \hat{K}^{1/2} v_i v_i^t \hat{K}^{1/2} \left(\hat{H}(c) Z_n - Z_n \right) + 2\mathrm{tr}\left(\hat{H}(c) \hat{K}^{1/2} v_i v_i^t \hat{K}^{1/2} \right), \end{aligned}$$

where $\hat{h}_i$ is the model average estimator of h_i. Clearly, $\sum_{i=1}^{n} v_i v_i^t = U^t M U$, so if all elements of the vector $h = (h_1, ..., h_n)^t$ are focused parameters, then the criterion function is modified to

$$\begin{aligned} \mathcal{W}(\hat{h}) &= \sum_{i=1}^{n} \mathcal{W}(\hat{h}_i) = \left(\hat{H}(c) Z_n - Z_n \right)^t \hat{K}^{1/2} \sum_{i=1}^{n} v_i v_i^t \hat{K}^{1/2} \left(\hat{H}(c) Z_n - Z_n \right) \\ &\quad + 2\mathrm{tr}\left(\hat{H}(c) \hat{K}^{1/2} \sum_{i=1}^{n} v_i v_i^t \hat{K}^{1/2} \right) \\ &= \left(\hat{H}(c) Z_n - Z_n \right)^t \hat{K}^{1/2} U^t M U \hat{K}^{1/2} \left(\hat{H}(c) Z_n - Z_n \right) \\ &\quad + 2\mathrm{tr}\left(\hat{H}(c) \hat{K}^{1/2} U^t M U \hat{K}^{1/2} \right) \\ &= n\hat{\sigma}^2 \left(\hat{H}(c) Z_n - Z_n \right)^t \left(\hat{H}(c) Z_n - Z_n \right) + 2n\hat{\sigma}^2 \mathrm{tr}\left(\hat{H}(c) \right) \\ &= n\hat{\sigma}^2 \left(\hat{H}(c) Z_n - Z_n \right)^t \left(\hat{H}(c) Z_n - Z_n \right) + 2n\hat{\sigma}^2 c^t g - 2n\hat{\sigma}^2 \tilde{p}, \end{aligned} \tag{3.3}$$

where $\hat{h} = (\hat{h}_1, ..., \hat{h}_n)^t$ and the vector g consists of the numbers of regressors of sub-models.

By (3.2) and some calculations, we have $\hat{H}_S = (U^t M U)^{1/2} \pi_S^t \left(\pi_S U^t M U \pi_S^t \right)^{-1} \pi_S (U^t M U)^{1/2}$ that equals W_i of [15], given that the ith sub-model in their paper is sub-model S (see Appendix A.1 for the detailed proof), and $Z_n = \hat{K}^{-1/2} D_n = (n\hat{\sigma}^2)^{-1/2} (U^t M U)^{1/2} \sqrt{n} \hat{\gamma}_{\mathrm{full}}$ that equals $\hat{\sigma}^{-1}\hat{\theta}$ of [15]. So by (A.14) of [15] and (3.3), we have

$$\mathcal{W}(\hat{h}) = n \left(\hat{h} - Y \right)^t \left(\hat{h} - Y \right) + 2n\hat{\sigma}^2 c^t g - n\tilde{p}\hat{\sigma}^2 - n^2\hat{\sigma}^2 + n\tilde{q}\hat{\sigma}^2,$$

where the last three terms are unrelated to weights. Hence, when the weights are non-random, the criterion function $\mathcal{W}(\hat{h})$ is equivalent to the Mallows criterion of [7].

Next, we consider the following random weights introduced by [15]

$$c(S|Z_n, \hat{\sigma}^2) = \frac{\bar{a}^{g_S} (n - g_S)^{\bar{b}} (\hat{\sigma}_S^2)^{\bar{c}}}{\sum\limits_{S'} \bar{a}^{g_{S'}} (n - g_{S'})^{\bar{b}} (\hat{\sigma}_{S'}^2)^{\bar{c}}},$$

where $\bar{a}(> 0)$, $\bar{b}(\geqslant 0)$, and $\bar{c}(\leqslant 0)$ are constants, g_S is the number of regressors in sub-model S, and $\hat{\sigma}_S^2$ is the maximum likelihood estimator of σ^2 based on sub-model S, which is a function of Z_n and $\hat{\sigma}^2$. So

in this case, $\hat{a}^*$ is also a function of $\hat{\sigma}^2$, and we rewrite it as $\hat{a}^*(Z_n, \hat{\sigma}^2)$. Noting that Z_n and $\hat{\sigma}^2$ are independent, from the derivations in (3.3), we see that under the random weights $c(S|Z_n, \hat{\sigma}^2)$'s,

$$\begin{aligned}\mathcal{W}(\hat{h}) &= n\left(\hat{h}-Y\right)^t\left(\hat{h}-Y\right)+2n\hat{\sigma}^2\mathrm{tr}\left(\frac{\partial\hat{a}^*(Z_n,\hat{\sigma}^2)}{\partial Z_n^t}\right)+n\tilde{p}\hat{\sigma}^2-n^2\hat{\sigma}^2+n\tilde{q}\hat{\sigma}^2\\ &= n\left(\hat{h}-Y\right)^t\left(\hat{h}-Y\right)+2n\hat{\sigma}^2\mathrm{tr}\left\{\hat{H}(c(Z_n,\hat{\sigma}^2))+\sum_S\frac{\partial c(S|Z_n,\hat{\sigma}^2)}{\partial Z_n}Z_n^t\hat{H}_S\right\}\\ &\quad +n\tilde{p}\hat{\sigma}^2-n^2\hat{\sigma}^2+n\tilde{q}\hat{\sigma}^2,\end{aligned}$$

where the last three terms are unrelated to weights and $\hat{H}(c(Z_n,\hat{\sigma}^2))+\sum_S\frac{\partial c(S|Z_n,\hat{\sigma}^2)}{\partial Z_n}Z_n^t\hat{H}_S$ is equal to $\Psi_1(\hat{\theta},\hat{\sigma}^2)$ of [15] with $Q=I_q$. So from (10) of [15] and the first step of (A.18) of [15], we obtain the equivalence between the OPT criterion and the criterion $\mathcal{W}(\hat{h})$.

4 Asymptotic optimality

Since the random weight vector $c(Z)$ in $\mathcal{W}^*(\hat{\mu})$ is an unspecified function of Z, the optimality using $c(Z)$ is hard to be defined if not impossible. Hence, in this section, we will show the asymptotic optimality of the FMA estimator with weights minimizing the criterion (2.7), written as $\mathcal{W}_0(c)$ with c denoting non-random weight vector. By (2.4) and some calculations, we see that the asymptotic risk of $\hat{\mu}(c)$ is given by

$$R_a(\hat{\mu})=\tau_0^2+\omega^tK^{1/2}H^2(c)K^{1/2}\omega+(\omega^tK^{1/2}L(c)a^*)^2\equiv R_a(c), \tag{4.1}$$

where $L(c)=I_q-H(c)$.

Let $\mathcal{C}=\left\{c\in[0,1]^{2^q}:\sum_S c_S=1\right\}$, a general weight set, and $\mathcal{C}^*$ be any subset of $\mathcal{C}$ including itself. Denote $\hat{c}=\underset{c\in\mathcal{C}^*}{\mathrm{argmin}}\mathcal{W}_0(c)$, the weight vector obtained by minimizing the criterion $\mathcal{W}_0(c)$ with c being restricted in the set $\mathcal{C}^*$. Let $\xi_n=\inf\limits_{c\in\mathcal{C}^*}R_a(c)$ and $\eta_n=\max\limits_S R_a(c_S^0)$ be the maximum risk based on a single sub-model, where c_S^0 is a weight vector with one for sub-model S and 0 otherwise. The following theorem builds the asymptotic optimality of $\hat{\mu}(\hat{c})$.

Theorem 4.1. *When $n\to\infty$, provided that the condition*

$$\xi_n^{-2}\eta_n\to 0 \tag{4.2}$$

is satisfied, then

$$R_a(\hat{c})\xi_n^{-1}\xrightarrow{p}1. \tag{4.3}$$

Proof: See the Appendix A.2.

Theorem 4.1 indicates that when the weight vector c is constrained to the set $\mathcal{C}^*$ and selected based on an auxiliary sample as $\hat{K}$ in [6], the asymptotic risk using the weight vector $\hat{c}$ is asymptotically equivalent to the infeasible optimal risk.

Remark 4.2. The condition (4.2) is similar to the condition (8) of [19] and the condition (22) of [15], where many theories and examples are used to show the rationality of these conditions. In [19], the weights are constrained to the set H_n that corresponds to the set $\mathcal{C}$ of the current paper. While [15] adopted the weight set ruling out the case where all models forming the model average are unbiased.

Remark 4.3. It is also worthy of pointing out that the asymptotic optimality properties shown in [7], [19] and [15] are on estimating means of responses, while our asymptotic optimality is on a focused parameter which can be any function of parameters in (2.1) given that it satisfies the condition in Lemma 3.3 of [9]. Another difference is that in the current paper we focus on asymptotic risk, while in those previous work, accurate risk or loss is studied.

5 Simulation study

We now conduct simulation experiments to evaluate the performance of the GOPT criterion given in (2.8) and use it to compare the well-known and commonly used weights: S-AIC, S-BIC, and S-FIC.

Example 5.1. We generate data from the linear regression model (3.1) with $x_i = (x_{i1}, x_{i2})^t$ following the distribution of $N(0, I_{2\times 2})$, $\beta = (1,2)^t$, $u_i = (u_{i1}, u_{i2}, u_{i3})^t$ following the distribution of $N(0, I_{3\times 3})$, $\gamma = \delta/\sqrt{n}$ with $\delta = (1,1,1)^t$, and $\varepsilon_i \sim N(0, \sigma^2)$, independent of x_i and u_i. The estimand is $\mu = x^t\beta + u^t\gamma$ with $x = (0.2, 0.3)^t$ and $u = (0.3, -0.1, 0.3)^t$.

Repeating B times, we obtain the average value of $\mathcal{W}(\hat{\mu})$

$$\mathcal{W} = \frac{1}{B}\sum_{b=1}^{B} \mathcal{W}^{(b)}(\hat{\mu})$$

for the S-AIC, S-BIC, and S-FIC weights, respectively, where $\mathcal{W}^{(b)}(\hat{\mu})$ is the value for the b-th run. In this and next sections, we let $\kappa = 1$ for S-FIC weights (Of course, we can also consider other values of κ). For evaluating the performance of the GOPT criterion in (2.8), we also calculate the risks of the estimators based on these three methods in estimating $\sqrt{n}\mu$. We define

$$\text{risk} = \frac{1}{B}\sum_{b=1}^{B} n(\mu - \hat{\mu}^{(b)})^2,$$

where $\hat{\mu}^{(b)}$ denotes the estimator for the b-th run and $B = 5000$. The sample size n is varied at $n = 100, 200$ and 400, and σ^2 is set as $0.1, 0.5, 1$ and 4. The results of this simulation are presented in Table 1.

We observe from Table 1 that the order of $\mathcal{W}_{\text{S-AIC}}$, $\mathcal{W}_{\text{S-BIC}}$ and $\mathcal{W}_{\text{S-FIC}}$ is exactly the same as that of $\text{risk}_{\text{S-AIC}}$, $\text{risk}_{\text{S-BIC}}$ and $\text{risk}_{\text{S-FIC}}$ in all cases, showing the good performance of the GOPT criterion for choosing weights. Taking the first case with $n = 100$ and $\sigma^2 = 0.1$ as an example, since $\mathcal{W}_{\text{S-FIC}} < \mathcal{W}_{\text{S-AIC}} < \mathcal{W}_{\text{S-BIC}}$, we should choose S-FIC weight, which is also the best by the real risk: $\text{risk}_{\text{S-FIC}} < \text{risk}_{\text{S-AIC}} < \text{risk}_{\text{S-BIC}}$.

Example 5.2. We generate data from the following logistic regression model

$$\Pr(Y_i = 1|x_i, u_i) = \frac{\exp(x_i^t\beta + u_i^t\gamma)}{1 + \exp(x_i^t\beta + u_i^t\gamma)}, \quad i = 1, ..., n,$$

where $x_i = (x_{i1}, x_{i2})^t$, $x_{i1} = 1$ is a constant, x_{i2} follows the distribution of $N(0,1)$, $\beta = (1,1)^t$, $u_i = (u_{i1}, u_{i2}, u_{i3})^t$ follows the distribution of $N(0, I_{3\times 3})$, and $\gamma = \delta/\sqrt{n}$ with $\delta = (1,1,1)^t$. We consider 10 estimands: $p_j(x_{0,j}, u_{0,j}) = \frac{\exp(x_{0,j}^t\beta + u_{0,j}^t\gamma)}{1+\exp(x_{0,j}^t\beta + u_{0,j}^t\gamma)}$, $j = 1, ..., 10$, where the components $x_{0,j,2}$, $u_{0,j,1}$, $u_{0,j,2}$, and $u_{0,j,3}$ are all generated from the normal distribution of $N(0,1)$ and $x_{0,j,1} = 1$.

From [4], we have $J_{n,\text{full}} = n^{-1}\sum_{i=1}^{n} p_i(1-p_i)\begin{pmatrix} x_ix_i^t & x_iu_i^t \\ u_ix_i^t & u_iu_i^t \end{pmatrix}$ with $p_i = \Pr(Y_i = 1|x_i, u_i)$, and $\omega_j = p_j(x_{0,j}, u_{0,j})\{1 - p_j(x_{0,j}, u_{0,j})\}(J_{n,10}J_{n,00}^{-1}x_{0,j} - u_{0,j})$. Like [4], in our simulation, we estimate the matrix $J_{n,\text{full}}$ using the full model, although we could also have estimated it from the narrow model with $\gamma = 0$. Correspondingly, $\hat{K}$ etc. can be obtained. We let the sample size n vary at $n = 100, 200$ and 400, and calculate the $\mathcal{W}$ and risk values for each estimand using the same way as in Example 1. The calculation results are summarized in Table A.2.

From Table A.2, it is seen that in 29 of all 30 cases, the selection results using the GOPT criterion and using the risk itself are the same. The exception happens when we estimate the 4-th estimand where $x_{0,j} = (1, 1.9790)^t$ and $u_{0,j} = (0.4160, 0.6112, 0.5953)^t$. For this estimand, when the sample size $n = 100$, the calculated risk suggest choosing S-FIC weight, but the GOPT criterion selects S-BIC. It is worthwhile mentioning that in this case, the $\mathcal{W}$ values with S-BIC and S-FIC are very close (which are 0.0302 and 0.0303, respectively). When the sample size is increased to 200 and 400, the choices by GOPT are in concord with those using risk.

6 Applications to real data

In this section, we apply the GOPT criterion of weight choice to the real 'survival' data and 'birth' data, and to assess the performances of the estimators based on the S-AIC, S-BIC, S-FIC and the full model. For the estimators from the full model, the standard errors (SEs) and confidence intervals (CIs) are calculated by utilizing the distribution of Λ_{full} (see (2.2)), and the corresponding $\mathcal{W}$ values are calculated using the expression (2.8) with the fixed weight where the element corresponding to the full model is one, and the others are zeros. As [9], the SEs and CIs for the random weight estimators are obtained by simulating the limit distribution of Λ given in (2.3) where the unknown parameters are also estimated under the full model.

6.1 Analysis of the survival data

The survival data from [11] consists of 54 records of patients. The response variable is *ltime*, the denary logarithm of the survival time of patient. The four covariates are *clot* (a blood clotting score), *prog* (a prognostic index, including age), *enz* (an enzyme function score) and *liv* (a liver function test score). Based on the previous study [11], the variable *liv* may not be necessary to be included in the model. Therefore, we consider the following linear model:

$$ltime = \theta_1 + \theta_2 clot + \theta_3 prog + \theta_4 enz + \gamma liv + \epsilon,$$

where ϵ is assumed to be $N(0, \sigma^2 I_n)$, and γ is an uncertain parameter with $\gamma_0 = 0$.

We apply our weight choice method to the estimation problem of coefficients, that is, we set the parameter of interest to be $\mu = \theta_i, i = 1, ..., 4$, and $\mu = \gamma$, respectively. By using different weights, we obtain the estimates along with the SEs, the $\mathcal{W}$ vaules, and the widths of the 95% CIs (see Table 3). Note that when there is only one uncertain covariate, the S-AIC and S-FIC give the same results (see also [9]), so no results are reported for the S-AIC. It can be observed that for each case, we have $\mathcal{W}_{\text{S-BIC}} < \mathcal{W}_{\text{S-FIC}} < \mathcal{W}_{\text{full model}}$, $\text{SE}_{\text{S-BIC}} < \text{SE}_{\text{S-FIC}} < \text{SE}_{\text{full model}}$, and $\text{Width}_{\text{S-BIC}} < \text{Width}_{\text{S-FIC}} < \text{Width}_{\text{full model}}$, indicating that for the estimation of each coefficient, choosing the weight with the smallest $\mathcal{W}$ value can lead to the estimator with the smallest SE and narrowest CI.

6.2 Analysis of the birth data

This example involves the data for a study of risk factors associated with the low infant birth weight. The data, collected at Baystate Medical Center, Springfield, Massachusets, during 1986 and presented in [13], contains the observations for 189 women, 59 of which had the low birth weight babies and 130 of which had the normal birth weight babies. Thus the response 'LOW' is a binary variable which indicates a low birth weight baby if it equals 1, and a normal birth weight baby if it equals 0. The nine covariates are respectively women's age (AGE), weight of the subject at her last menstrual period (LWT), $RACE\_2$ (which equals 1 if the race of woman is black, and 0 otherwise), $RACE\_3$ (which equals 1 if the race is neither black nor white, and 0 otherwise), the number of physician visits during the first trimester of the pregnancy (FTV), *smoke* (which equals 1 if smokes during pregnancy, and 0 otherwise), *ptl* (which equals 1 if having a history of premature labor, and 0 otherwise), *ht* (which equals 1 if having a history of hypertension, and 0 otherwise), and *ui* (which equals 1 if presents uterine irritability, and 0 otherwise). Previous analysis of these data (see, for example, [13]) preferred to only the first five variables in the model fitting. Thus, we consider the following logistic model

$$\begin{aligned} logit \Pr(LOW = 1) &= \theta_1 + \theta_2 AGE + \theta_3 LWT + \theta_4 RACE\_2 + \theta_5 RACE\_3 + \theta_6 FTV \\ &\quad + \gamma_1 smoke + \gamma_2 ptl + \gamma_3 ht + \gamma_4 ui \end{aligned}$$

with $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)'$ being an uncertain parameter vector and $\gamma_0 = 0_{4\times 1}$. Analogous to the analysis for the survival data, we first consider the estimation of the coefficients. The results are presented in Table 4.

Observing these results, we find that for almost all the cases, the smallest SEs and narrowest CIs can be obtained by choosing the smallest $\mathcal{W}$ values. In detail, for estimating all ten coefficients, the estimators with the narrowest CIs can be obtained by choosing the weights with the smallest $\mathcal{W}$ values; for estimating nine of them, the estimators with the smallest SEs can be obtained by choosing the weights with the smallest $\mathcal{W}$ values. The exception happens in estimating the coefficient of *LWT*, where $\mathcal{W}_{\text{S-BIC}}$ is the minimum among the four $\mathcal{W}$ values, but the SE of the estimator with S-FIC weight is the minimum among the four SEs.

We next take an interest in estimating the probabilities of the low birth weight for the average "white", "black" and "other" mothers. The similar analysis has been done by [9]. Our results are provided in Table 5. From the table, we see that for each case, choosing the weights with the smallest $\mathcal{W}$ values leads to the estimators with the smallest SEs. For the CI, when estimating the probability of the low birth weight for the average "white" mother, we get the estimator with the narrowest CI by choosing the weight with the smallest $\mathcal{W}$ value; in the two other cases, however, the estimators by the full model have the narrowest CIs, and our weight choice method chooses the S-BIC and S-AIC estimators, respectively.

7 Choice between model averaging and model selection

Model averaging and model selection are both feasible when the number of candidate models is larger than 1. Although the advantages of model averaging have been demonstrated by theories, simulations and real data examples in the literature (see, for example , [7, 14, 24]), the advantages should not be overplayed. [8] showed situations where model averaging results in outcomes far worse than those from the full model, and [24] also presented some cases where model selection does perform better than model averaging. Thus, when should practitioners choose model averaging?

Some pioneering work has been done for this troublesome problem. Examples are [23] and [16], where three kinds of model selection diagnostics were introduced including the perturbation instability in estimation (PIE), bootstrap instability in selection (BIS) and bootstrap instability in estimation (BIE). These diagnostics were used to measure the model selection instability (MSI) or model selection instability in estimation (MSIE). When these diagnostics indicate that there do not exist high MSI and MSIE, estimation based on a properly selected model is sound; otherwise, model averaging is preferred.

Here, by making use of the GOPT criterion, we introduce a method for choice between model averaging and model selection. Take the model averaging by S-FIC and model selection by FIC (hereafter, they are denoted as S-FIC and FIC, respectively) as an example. From (2.11), we see that when $\kappa \to \infty$, the weight vector $c_{\text{FIC}}(Z_n)$ tends to a special weight vector with the elements of one and zeros almost surely (unless there are equal FIC values, which hardly occurs in practice), denoted as c_{FIC_0}. This means that $\kappa \to \infty$ corresponds to the model selection. We rewrite $\mathcal{W}_{\text{S-FIC}}$ as

$$\begin{aligned}\mathcal{W}_{\text{S-FIC}}(\kappa) &= \left\{\hat{\omega}^t\hat{K}^{1/2}(\hat{H}(c_{\text{FIC}}(Z_n)) - I_q)Z_n\right\}^2 \\ &\quad +2\hat{\omega}^t\hat{K}^{1/2}\hat{H}(c_{\text{FIC}}(Z_n))\hat{K}^{1/2}\hat{\omega} + 2\kappa\Delta\left(c_{\text{FIC}}(Z_n)\right),\end{aligned}$$

where

$$\begin{aligned}\Delta\left(c_{\text{FIC}}(Z_n)\right) &= -\hat{\omega}^t\hat{K}^{1/2}\left(\sum_S c_{\text{FIC}}(S \mid Z_n)\hat{H}_S Z_n A_S\right)\hat{K}^{1/2}\hat{\omega} \\ &\quad +\hat{\omega}^t\hat{K}^{1/2}\left(\hat{H}(c_{\text{FIC}}(Z_n))Z_n\sum_S c_{\text{FIC}}(S \mid Z_n)A_S\right)\hat{K}^{1/2}\hat{\omega}.\end{aligned}$$

When $\kappa \to \infty$, it is straightforward to prove that $\kappa\Delta\left(c_{\text{FIC}}(Z_n)\right) \to 0$ almost surely, and thus $\mathcal{W}_{\text{S-FIC}}(\kappa) \to \left\{\hat{\omega}^t\hat{K}^{1/2}(\hat{H}(c_{\text{FIC}_0}) - I_q)Z_n\right\}^2 + 2\hat{\omega}^t\hat{K}^{1/2}\hat{H}(c_{\text{FIC}_0})\hat{K}^{1/2}\hat{\omega} \equiv \mathcal{W}_{\text{FIC}}$ almost surely. The commonly used S-FIC corresponds to $\kappa = 1$. If $\mathcal{W}_{\text{S-FIC}}(\kappa = 1) < \mathcal{W}_{\text{FIC}}$, we choose S-FIC; otherwise, we choose FIC. In other

words, the comparison of values of $\mathcal{W}_{\text{S-FIC}}(\kappa = 1)$ and $\mathcal{W}_{\text{FIC}}$ can be served as choosing between S-FIC and FIC. After replacing 1/2 by κ in (2.9) and (2.10), it is straightforward to develop similar methods for choosing between S-AIC and AIC or S-BIC and BIC.

Now, we conduct a simple simulation to see the performance of the method. We generate data from the linear regression model with $\varepsilon_i \sim N(0,1)$, $(x_{i1}, u_{i1}, u_{i2})^t \sim N(0, I_{3\times 3})$, $\beta = 0.2$, $\delta = (0.1, 1)^t$, and take $n = 200$. The estimand is $\mu = 0.05\beta + (-1, 0.1)\gamma$. We run 10000 replications. In each replication, the squared losses for estimating μ by S-FIC and FIC are compared. If S-FIC (or FIC) is chosen by our method and the corresponding squared loss is smaller, then our method makes a correct choice. The simulation shows that the correct proportion is 76.14%, which should be acceptable, as $\mathcal{W}_{\text{S-FIC}}(\kappa)$ is developed from asymptotic risk and some unknown quantities in $\mathcal{W}_{\text{S-FIC}}(\kappa)$ are directly replaced by their sample versions.

We further apply this method to the choice of FIC and S-FIC in the analysis of birth data. There are 189 observations in this data set with "id"s from 4 to 226. We use the first 100 observations with "id"s from 4 to 128 as a training sample and the last 89 observations as a testing sample. The S-FIC and FIC are performed and the inequality $\mathcal{W}_{\text{S-FIC}}(\kappa = 1) < \mathcal{W}_{\text{FIC}}$ is used to choose between them. The resultant correct proportion for all the 89 testing observations is 69.66%.

8 Concluding remarks

In this paper, we have studied the GOPT criterion of weight choice proposed by [15]. More specifically, the criterion has been connected to the Mallows and OPT criteria, and used to compare some existing weights: S-AIC, S-BIC and S-FIC weights. Simulation results show that the GOPT criterion can usually choose the best weight in the sense of minimizing the risk of estimation. Further, the real data analysis also suggests that the criterion is promising. In addition, an asymptotic optimality on the criterion in the case of non-random weights is proved and a method for the choice between model selection and model averaging is proposed on the basis of the GOPT criterion.

It should be noted that the criterion for weight choice is based on parametric models. Developing a criterion of choosing weights based on non-parametric models is no doubt an interesting topic and this warrants our further research. Moreover, if model selection or model average estimator is chosen by our proposed methods, the resultant estimator will end up being conditional on this choice and likely exhibit extremely complicated properties. A fruitful avenue for further research would be the derivation of the resultant estimator's full distribution. The recent related work of [17] may serve as a useful guide in this regard.

Acknowledgements Zhang's research was partially supported by the two grants 71101141 and 70933003 from the National Natural Science Foundation of China, Science Foundation of the Chinese Academy of Sciences, and NCMIS; Zou's research was partially supported by the two grants 70625004 and 11021161 from the National Natural Science Foundation of China, and the Hundred Talents Program of the Chinese Academy of Sciences; and Liang's research was partially supported by the NSF grant DMS-1007167.

Appendix

A.1 Proof of equivalence between H_S and W_i in [15]

Proof. First, we reorder the uncertain parameters such that $(\pi_{SC}^t, \pi_S^t) = I_q$. Let $V^tV = (U^tMU)^{-1}$ and partition V as $V = (V_1 : V_2)$ such that V_1 includes $q - |S|$ columns and V_2 the others, where $|S|$ is the number of the elements in S. Denote $J = I_q - V_1(V_1^tV_1)^{-1}V_1^t$ and $Q = (V_1^tV_1)^{-1}V_1^tV_2(V_2^tJV_2)^{-1/2}$. Then

we have

$$\begin{aligned}
&(U^tMU)^{-1/2}\left\{I_q-(U^tMU)^{-1/2}\pi_{S^C}^t\left(\pi_{S^C}(U^tMU)^{-1}\pi_{S^C}^t\right)^{-1}\pi_{S^C}(U^tMU)^{-1/2}\right\}\\
&\quad\times(U^tMU)^{-1/2}\\
&=(U^tMU)^{-1}-(U^tMU)^{-1}\pi_{S^C}^t\left(\pi_{S^C}(U^tMU)^{-1}\pi_{S^C}^t\right)^{-1}\pi_{S^C}(U^tMU)^{-1}\\
&=(U^tMU)^{-1}-(U^tMU)^{-1}\begin{pmatrix}(V_1^tV_1)^{-1}&0\\0&0\end{pmatrix}(U^tMU)^{-1}\\
&=\begin{pmatrix}0&0\\0&V_2^tJV_2\end{pmatrix}.
\end{aligned}\tag{A.1}$$

It is well-known [18] that

$$U^tMU=\begin{pmatrix}V_1^tV_1&V_1^tV_2\\V_2^tV_1&V_2^tV_2\end{pmatrix}^{-1}=\begin{pmatrix}(V_1^tV_1)^{-1}+QQ^t&-Q(V_2^tJV_2)^{-1/2}\\-(V_2^tJV_2)^{-1/2}Q^t&(V_2^tJV_2)^{-1}\end{pmatrix},$$

by which we obtain

$$\begin{aligned}
&(U^tMU)^{-1/2}H_S(U^tMU)^{-1/2}\\
&=(U^tMU)^{-1/2}(U^tMU)^{1/2}\pi_S^t\left(\pi_SU^tMU\pi_S^t\right)^{-1}\pi_S(U^tMU)^{1/2}(U^tMU)^{-1/2}\\
&=\begin{pmatrix}0&0\\0&V_2^tJV_2\end{pmatrix}.
\end{aligned}\tag{A.2}$$

Now, by comparing (A.1) and (A.2), the conclusion is confirmed. □

A.2 Proof of Theorem 1

Proof. From the definition of infimum, there exists a series of non-negative ϑ_n and vectors $c(n)\in\mathcal{C}^*$ such that $\vartheta_n\to 0$ and

$$\xi_n=R_a(c(n))-\vartheta_n,\tag{A.3}$$

as $n\to\infty$. Define

$$t_n(c)=\mathcal{W}_0(c)+\tau_0^2-\omega^tK\omega-R_a(c).\tag{A.4}$$

For any $\delta>0$, we have

$$\begin{aligned}
&\Pr\left\{\left|\frac{\xi_n}{R_a(\hat{c})}-1\right|>\delta\right\}=\Pr\left\{\frac{R_a(\hat{c})-\xi_n}{R_a(\hat{c})}>\delta\right\}\\
&=\Pr\left\{\frac{\inf_{c\in\mathcal{C}^*}\left(R_a(c)+t_n(c)\right)-t_n(\hat{c})-\xi_n}{R_a(\hat{c})}>\delta\right\}\\
&\leqslant\Pr\left\{\xi_n^{-1}\left[R_a(c(n))+t_n(c(n))-t_n(\hat{c})-R_a(c(n))+\vartheta_n\right]>\delta\right\}\\
&=\Pr\left\{\xi_n^{-1}\left[t_n(c(n))-t_n(\hat{c})+\vartheta_n\right]>\delta\right\}.
\end{aligned}\tag{A.5}$$

Thus, by (4.2) and $\vartheta_n\to 0$, we need only to prove that, as $n\to\infty$,

$$\xi_n^{-1}\sup_{c\in\mathcal{C}^*}|t_n(c)|\xrightarrow{p}0.\tag{A.6}$$

Let $\varpi=Z-a^*$, and then $\varpi\sim N(0,I_q)$. By (2.7), (4.1) and (A.4), we have

$$t_n(c)=\left(\omega^tK^{1/2}L(c)\varpi\right)^2-\omega^tK^{1/2}L^2(c)K^{1/2}\omega+2\omega^tK^{1/2}L(c)a^*\omega^tK^{1/2}L(c)\varpi.\tag{A.7}$$

We observe, for any $\delta > 0$, that

$$\begin{aligned}
&\Pr\left\{\xi_n^{-1}\sup_{c\in\mathcal{C}^*}|t_n(c)|>\delta\right\} \leqslant \Pr\left\{\xi_n^{-1}\sup_{c\in\mathcal{C}^*}\left|\left(\omega^tK^{1/2}L(c)\varpi\right)^2-\omega^tK^{1/2}L^2(c)K^{1/2}\omega\right|\right.\\
&\qquad\left.+\xi_n^{-1}\sup_{c\in\mathcal{C}^*}\left|2\omega^tK^{1/2}L(c)a^*\omega^tK^{1/2}L(c)\varpi\right|>\delta\right\}\\
&\leqslant \Pr\left\{\xi_n^{-1}\sup_{c\in\mathcal{C}}\left|\left(\omega^tK^{1/2}L(c)\varpi\right)^2-\omega^tK^{1/2}L^2(c)K^{1/2}\omega\right|\right.\\
&\qquad\left.+\xi_n^{-1}\sup_{c\in\mathcal{C}}\left|2\omega^tK^{1/2}L(c)a^*\omega^tK^{1/2}L(c)\varpi\right|>\delta\right\}\\
&\leqslant \sum_i\sum_j\Pr\left\{\xi_n^{-1}\left|\varpi^tL(c_i^0)K^{1/2}\omega\omega^tK^{1/2}L(c_j^0)\varpi-\omega^tK^{1/2}L(c_i^0)L(c_j^0)K^{1/2}\omega\right|>\delta/2\right\}\\
&\qquad+\sum_i\sum_j\Pr\left\{\xi_n^{-1}\left|2\omega^tK^{1/2}L(c_i^0)a^*\omega^tK^{1/2}L(c_j^0)\varpi\right|>\delta/2\right\}\\
&\leqslant 4\xi_n^{-2}\delta^{-2}C\sum_i\sum_j\mathrm{tr}\left(L(c_i^0)K^{1/2}\omega\omega^tK^{1/2}L(c_j^0)L(c_j^0)K^{1/2}\omega\omega^tK^{1/2}L(c_i^0)\right)\\
&\qquad+16\xi_n^{-2}\delta^{-2}\sum_i\sum_j\left(\omega^tK^{1/2}L(c_i^0)a^*\right)^2\omega^tK^{1/2}L(c_j^0)L(c_j^0)K^{1/2}\omega\\
&\leqslant 4\xi_n^{-2}\delta^{-2}C\sum_i\sum_j\omega^tK^{1/2}L(c_j^0)L(c_j^0)K^{1/2}\omega\omega^tK^{1/2}L(c_i^0)L(c_i^0)K^{1/2}\omega\\
&\qquad+16\xi_n^{-2}\eta_n\delta^{-2}\sum_i\sum_j\omega^tK^{1/2}L(c_j^0)L(c_j^0)K^{1/2}\omega, \qquad (A.8)
\end{aligned}$$

where C is a positive constant, the fourth inequality is from Theorem 2 of [21], and the last inequality is from (4.1). Now, by (A.8), the condition (4.2), and the fact that J_{full} is invertible and unrelated to n, the result (A.6) is obtained. Thus, the proof is completed. □

References

1 Bates J M, Granger C W J. The combination of forecasts. Operations Research Quarterly, 1969, 20: 451–468
2 Buckland S T, Burnham K P, Augustin N H. Model selection: An integral part of inference. Biometrics, 1997, 53: 603–618
3 Claeskens G, Croux C, Venkerckhoven J. Variable selection for logit regression using a prediction-focused information criterion. Biometrics, 2006, 62: 972–979
4 Claeskens G, Hjort N L. The focused information criterion. Journal of the American Statistical Association, 2003, 98: 900–945
5 Claeskens G, Hjort N L. Model Selection and Model Averaging. Cambridge: Cambridge University Press, 2008
6 Donald S, Newey W. Choosing the number of instruments. Econometrica, 2001, 69: 1161–1191
7 Hansen B E. Least squares model averaging. Econometrica, 2007, 75: 1175–1189
8 Hendry D F, Reade J J. Forecasting using model averaging. Technical report, Nuffield College, University of Oxford, Oxford UK, 2011
9 Hjort N L, Claeskens G. Frequentist model average estimators. Journal of the American Statistical Association, 2003, 98: 879–899
10 Hjort N L, Claeskens G. Focused information criteria and model averaging for the Cox hazard regression model. Journal of the American Statistical Association, 2006, 110: 1449–1464
11 Hocking R R. Methods and Applications of Linear Models. Second ed, Hoboken NJ: John Wiley & Sons, 2003
12 Hoeting J A, Madigan D, Raftery A E, Volinsky C T. Bayesian model averaging: A tutorial. Statistical Science, 1999, 14: 382–417
13 Hosmer D W, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons, 1989
14 Leung G, Barron A R. Infromation theory and mixing least-squares regressions. IEEE Transactions on Information Theory, 2006, 52: 3396–3410

15 Liang H, Zou G H, Wan A T K, Zhang X Y. Optimal weight choice for frequentist model average estimators. Journal of the American Statistical Association, 2011, 106: 1053–1066
16 Liu S, Yang Y H. Combining models in longitudinal data analysis. Annals of the Institute of Statistical Mathematics, 2011, 64: 233–254
17 Pötscher B M. The distribution of model averaging estimators and an impossibility result regarding its estimation IMS Lecture Notes-Monograph Series: Time Series and Related Topics, 2006, 52: 113–129
18 Rao C R. Linear Statistical Inference and Its Applications. Second ed, John Wiley & Sons, 1973
19 Wan A T K, Zhang X Y, Zou G H. Least squares model combining by Mallows criterion. Journal of Econometrics, 2010, 156: 277–283
20 Wang H, Zhang X Y, Zou G H. Frequentist model averaging estimation: A review. Journal of Systems Science and Complexity, 2009, 22: 732–748
21 Whittle P. Bounds for the moments of linear and quadratic forms in independent variables. Theory of Probability and Its Applications, 1960, 5: 302–305
22 Yang Y H. Adaptive regression by mixing. Journal of the American Statistical Association, 2001, 96: 574–586
23 Yuan Z, Yang Y H. Combining linear regression models: When and how? Journal of the American Statistical Association, 2005, 100: 1202–1214
24 Zhang X Y, Liang H. Focused information criterion and model averaging for generalized additive partial linear models. Annals of Statistics, 2011, 39: 174–200

Table 1 Simulation study for Example 1

		$\mathcal{W}$			risk		
n	σ^2	S-AIC	S-BIC	S-FIC	S-AIC	S-BIC	S-FIC
100	0.1	0.0488	0.0619	0.0469	0.0430	0.0562	0.0410
	0.5	0.2118	0.2399	0.1926	0.1862	0.2138	0.1653
	1.0	0.3640	0.3774	0.3396	0.3088	0.3203	0.2842
	4.0	1.1921	1.0859	1.1573	0.9524	0.8312	0.9142
200	0.1	0.0476	0.0649	0.0463	0.0417	0.0588	0.0402
	0.5	0.2097	0.2486	0.1892	0.1772	0.2187	0.1576
	1.0	0.3586	0.3802	0.3345	0.2987	0.3158	0.2734
	4.0	1.1746	1.0516	1.1349	0.9431	0.8097	0.9004
400	0.1	0.0471	0.0696	0.0459	0.0421	0.0652	0.0411
	0.5	0.2087	0.2573	0.1881	0.1766	0.2313	0.1569
	1.0	0.3572	0.3868	0.3336	0.2992	0.3278	0.2729
	4.0	1.1682	1.0391	1.1324	0.9423	0.7962	0.9062

Table 2 Simulation study for Example 2

		$\mathcal{W}$			risk		
$(x_{0,j}^t, u_{0,j}^t)$	n	S-AIC	S-BIC	S-FIC	S-AIC	S-BIC	S-FIC
(1,-0.8186; 0.2621,-0.1829,0.5601)	100	0.2799	0.2532	0.2730	0.6133	0.5670	0.6125
	200	0.2457	0.2175	0.2400	0.5795	0.5347	0.5769
	400	0.2259	0.1930	0.2218	0.5664	0.5335	0.5625
(1,1.2269; -1.7830,1.0041,-0.2150)	100	0.4087	0.3556	0.3895	0.3052	0.2375	0.2869
	200	0.3487	0.2953	0.3320	0.2934	0.2266	0.2792
	400	0.3134	0.2593	0.3005	0.2937	0.2226	0.2815
(1,-0.0989; 0.9522,1.0264,0.3596)	100	0.6812	0.6020	0.6671	0.4698	0.4127	0.4675
	200	0.6957	0.6238	0.6820	0.4763	0.4212	0.4670
	400	0.7118	0.6449	0.6978	0.4879	0.4315	0.4778
(1,1.9790; 0.4160,0.6112,0.5953)	100	0.0308	0.0302	0.0303	0.1033	0.1114	0.1021
	200	0.0250	0.0251	0.0245	0.1022	0.1093	0.1010
	400	0.0213	0.0218	0.0209	0.0976	0.1022	0.0970
(1,-0.9277; 0.3522,0.6183,1.5049)	100	1.4362	1.2695	1.4193	1.1139	0.9747	1.1112
	200	1.4467	1.2768	1.4292	1.0555	0.9155	1.0468
	400	1.4374	1.2763	1.4262	1.0597	0.9312	1.0487
(1,0.3520; -0.8006,0.1797,0.1366)	100	0.1747	0.1478	0.1646	0.3067	0.2734	0.3026
	200	0.1589	0.1283	0.1507	0.2958	0.2653	0.2914
	400	0.1539	0.1215	0.1476	0.3010	0.2649	0.2961
(1,-1.4491; 1.2913,-1.4540,-0.3228)	100	1.6998	1.3823	1.6292	1.4588	1.1535	1.4139
	200	1.7352	1.3700	1.6630	1.4780	1.1245	1.4125
	400	1.7449	1.3501	1.6775	1.4625	1.0881	1.3922
(1,0.4255; -0.5472,0.9250,-0.4314)	100	0.2775	0.2300	0.2733	0.3199	0.2711	0.3152
	200	0.2673	0.2145	0.2614	0.3111	0.2621	0.3065
	400	0.2648	0.2084	0.2580	0.3207	0.2608	0.3133
(1,0.8213; 0.8870,0.1838,1.3619)	100	0.2731	0.2323	0.2681	0.2754	0.2617	0.2731
	200	0.2839	0.2441	0.2783	0.2945	0.2788	0.2896
	400	0.2901	0.2520	0.2858	0.2939	0.2758	0.2897
(1,0.3049; 0.3376,-0.8751,0.9054)	100	0.4073	0.3415	0.3914	0.3676	0.2931	0.3576
	200	0.3979	0.3244	0.3817	0.3719	0.2910	0.3564
	400	0.3879	0.3052	0.3733	0.3518	0.2742	0.3348

Table 3 Coefficient estimation for the survival data

	S-BIC	S-FIC	full model		S-BIC	S-FIC	full model
		Inter				*enz*	
Est.	0.4842459	0.4850180	0.4887377	Est.	0.0095173	0.0095099	0.0094744
SE	0.0443387	0.0457201	0.0502395	SE	0.0003291	0.0003447	0.0003963
Width	0.1742615	0.1793604	0.1969351	Width	0.0012863	0.0013544	0.0015535
$\mathcal{W}$	0.0102190	0.0210688	0.0724394	$\mathcal{W}$	0.0000009	0.0000019	0.0000066
		clot				*liv*	
Est.	0.0691431	0.0690371	0.0685265	Est.	0.0002337	0.0005232	0.0019180
SE	0.0044329	0.0046754	0.0054405	SE	0.0044107	0.0060014	0.0097108
Width	0.0173779	0.0182477	0.0213262	Width	0.0183241	0.0269679	0.0380658
$\mathcal{W}$	0.0001926	0.0003970	0.0013650	$\mathcal{W}$	0.0014367	0.0029621	0.0101844
		prog					
Est.	0.0092897	0.0092836	0.0092543				
SE	0.0003983	0.0004078	0.0004368				
Width	0.0015673	0.0016090	0.0017122				
$\mathcal{W}$	0.0000006	0.0000013	0.0000045				

Table 4 Coefficient estimation for the birth data

	S-AIC	S-BIC	S-FIC	full model		S-AIC	S-BIC	S-FIC	full model
		Inter					*FIV*		
Est.	0.5147	0.5432	0.5539	0.3524	Est.	-0.0025	-0.0231	0.0046	0.0638
SE	1.1597	1.1497	1.1677	1.1932	SE	0.1710	0.1703	0.1712	0.1718
Width	4.5494	4.4907	4.5883	4.6774	Width	0.6779	0.6720	0.6760	0.6736
$\mathcal{W}$	33.4671	21.3322	60.3931	68.9031	$\mathcal{W}$	0.1787	0.1558	0.3079	0.3031
		AGE					*smoke*		
Est.	-0.0254	-0.0232	-0.0241	-0.0273	Est.	0.2484	0.0646	0.2616	0.9403
SE	0.0374	0.0377	0.0374	0.0370	SE	0.2501	0.1535	0.2460	0.4020
Width	0.1468	0.1476	0.1462	0.1451	Width	1.1294	0.5672	1.0914	1.5756
$\mathcal{W}$	0.0180	0.0299	0.0219	0.0135	$\mathcal{W}$	17.5077	5.2426	16.8974	61.0729
		LWT					*ptl*		
Est.	-0.0119	-0.0111	-0.0125	-0.0149	Est.	0.6300	0.5392	0.6285	0.5630
SE	0.0066	0.0066	0.0065	0.0069	SE	0.4216	0.4934	0.4439	0.3456
Width	0.0259	0.0256	0.0256	0.0270	Width	1.5266	1.6055	1.5938	1.3546
$\mathcal{W}$	0.0015	0.0010	0.0012	0.0023	$\mathcal{W}$	60.6756	102.6005	65.2857	45.1420
		RACE_2					*ht*		
Est.	1.1143	1.0708	1.1216	1.2935	Est.	0.6034	0.1623	0.5637	1.8385
SE	0.5138	0.5100	0.5150	0.5274	SE	0.5081	0.3635	0.4747	0.6961
Width	2.0001	1.9982	2.0028	2.0675	Width	2.1242	1.5264	2.0119	2.7285
$\mathcal{W}$	2.1605	1.2165	4.1718	6.0321	$\mathcal{W}$	114.7926	76.0977	113.3889	183.1430
		RACE_3					*ui*		
Est.	0.5772	0.4991	0.6080	0.9097	Est.	0.6830	0.6887	0.7697	0.7403
SE	0.4084	0.3939	0.4086	0.4406	SE	0.5031	0.5821	0.5004	0.4595
Width	1.5842	1.5358	1.5796	1.7270	Width	2.0646	2.4368	2.0544	1.8010
$\mathcal{W}$	5.0888	1.6986	5.4254	17.5772	$\mathcal{W}$	81.3017	86.9877	80.6182	79.7962

Table 5 Estimation of the probability of low birth weight

	S-AIC	S-BIC	S-FIC	full model
		white		
Est.	0.21610	0.22155	0.21344	0.19661
SE	0.04345	0.04312	0.04346	0.04422
Width	0.17237	0.17047	0.17211	0.17370
$\mathcal{W}$	0.03103	0.02915	0.05709	0.06114
		black		
Est.	0.42031	0.42049	0.41709	0.41698
SE	0.10177	0.10176	0.10176	0.10179
Width	0.39778	0.39804	0.39779	0.37962
$\mathcal{W}$	0.00091	0.00059	0.00111	0.00172
		other		
Est.	0.36165	0.36327	0.36341	0.36138
SE	0.06304	0.06305	0.06304	0.06304
Width	0.24595	0.24565	0.24632	0.24265
$\mathcal{W}$	0.00326	0.00406	0.00390	0.00333