SCIENCE CHINA
Mathematics

Special Topic on Regression Model Inference
• ARTICLES •

February 2019 Vol. 62 No. 2: 205–226
https://doi.org/10.1007/s11425-018-9403-x

A general framework for frequentist model averaging

In Celebration of Professor Lincheng Zhao's 75th Birthday

Priyam Mitra[1], Heng Lian[2], Ritwik Mitra[3], Hua Liang[4] & Min-ge Xie[1,*]

[1]*Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ* 08854*, USA;*
[2]*Department of Mathematics, City University of Hong Kong, Hong Kong, China;*
[3]*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ* 08540*, USA;*
[4]*Department of Statistics, George Washington University, Washington, DC* 20052*, USA*

Email: pmitra@scarletmail.rutgers.edu, henglian@cityu.edu.hk, ritwikm@scarletmail.rutgers.edu, hliang@gwu.edu, mxie@stat.rutgers.edu

Received March 1, 2018; accepted September 22, 2018; published online January 16, 2019

Abstract Model selection strategies have been routinely employed to determine a model for data analysis in statistics, and further study and inference then often proceed as though the selected model were the true model that were known a priori. Model averaging approaches, on the other hand, try to combine estimators for a set of candidate models. Specifically, instead of deciding which model is the 'right' one, a model averaging approach suggests to fit a set of candidate models and average over the estimators using data adaptive weights. In this paper we establish a general frequentist model averaging framework that does not set any restrictions on the set of candidate models. It broadens the scope of the existing methodologies under the frequentist model averaging development. Assuming the data is from an unknown model, we derive the model averaging estimator and study its limiting distributions and related predictions while taking possible modeling biases into account. We propose a set of optimal weights to combine the individual estimators so that the expected mean squared error of the average estimator is minimized. Simulation studies are conducted to compare the performance of the estimator with that of the existing methods. The results show the benefits of the proposed approach over traditional model selection approaches as well as existing model averaging methods.

Keywords asymptotic distribution, bias variance trade-off, local mis-specification, model averaging estimators, optimal weight selection

MSC(2010) 62F99, 62J99

Citation: Mitra P, Lian H, Mitra R, et al. A General framework for frequentist model averaging. Sci China Math, 2019, 62: 205–226, https://doi.org/10.1007/s11425-018-9403-x

1 Introduction

When there are several plausible models to choose from but no definite scientific rationale to dictate which one should be used, a model selection method has been used traditionally to determine a 'correct' model for data analysis. Commonly used model selection methods, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), stepwise regression, best subset selection, penalised regression, etc., are data driven and different methods may use different criteria (see, e.g., [10] and the reference therein). Once a model is chosen, further analysis proceeds as if the model selected is the true one. This practice

* Corresponding author

does not account for the uncertainty introduced in the process due to model selection, and can often lead to faulty inference as discussed in [2,6,22], among others. Model averaging methods have been introduced to incorporate different models during analysis (see, e.g., [3]). Instead of deciding which model is the 'correct' one, a model averaging method uses a set of plausible candidate models. The candidate models are combined using some data-dependent weights to reflect the degree to which each candidate model is trusted.

Our research on model averaging is motivated in part by a real life example on a prostate cancer study where the relationship between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy were investigated. The variables included in the study are log cancer volume, log prostate weight, age, log of the amount of benign prostatic hyperplasia, seminal vesicle invasion, log of capsular penetration, Gleason score, and percent of Gleason scores 4 or 5. In analysis of such data, a common theme is that different model selection methods may choose different models as the 'true' one. For example, AIC and BIC, two commonly used model selection criteria, may pick two different models, as the criteria for selection are different. Such situations would certainly raise many questions in practice. For example, if the estimator is selected by using a model selection criterion, how would we address the possibility that the selection is a wrong model? Also, if different model selection methods give us different results, we might wonder how trustworthy the model selection procedures are. Instead of choosing one model using a model selection scheme, we can use an average of estimators from different models.

In [11], Hjort and Claeskens provided a formal theoretical treatment of frequentist model averaging approaches, which provided an in-depth understanding of the approaches. However, the development in [11] had an assumption that any extra parameters not in its defined "narrow model" will shrink to zero at an $\mathcal{O}(1/\sqrt{n})$ rate. This assumption essentially requires that all candidate models are within an $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. Although this assumption avoids a technical difficulty of handling biased estimators, in reality we do not know the true model and thus excluding from consideration those models that are beyond $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model appears to be very restrictive in practice. In this paper, we remove this restrictive assumption in [11] and develop frequentist model averaging approaches under a more general framework. Our model averaging scheme allows us to use all the potential candidate models available, even the ones that produce biased estimates.

The development is motivated by the familiar bias-variance trade-off. If we use an overly simple model, the parameter estimates will often be biased, but it can also possibly have less variance, because there are fewer parameters to estimate. Similarly, if a bigger model is used, the parameter estimates often have low or no bias but increased variance. It is possible that biased estimators may end up having lower mean squared errors than the bigger model or even the true model, and vice versa. In our development, we study the delicate balance between bias and variance in all possible models and utilize the knowledge to develop new frequentist model averaging approaches.

A key element of a model averaging method is selection of weights that help us build a combined model averaging estimator. The weights proposed in our development are based on the aforementioned bias-variance trade-off, anchoring on the mean squared error (MSE) of the overall model averaging estimator. The weighing scheme is similar to but not the same as that discussed in [20], in which the authors only focused on Gaussian linear regression models. Specifically, a consistent estimate of the mean squared error of the model averaging estimator is proposed, and the weights are chosen such that the MSE estimate is minimized.

A model averaging estimator combines a set of competing candidate models rather than choosing just one. It also provides an insurance against selecting a poor model thus improving the risk in estimation. In [3, 12], variable selection methods for the Cox proportional hazards regression model were discussed along with the choice of weights. In [9] a new set of weights was derived using Mallow's criterion. In [20], Liang et al. proposed an unbiased estimator of the risk and a set of optimal weights was chosen by minimizing the trace of the unbiased estimator. Further details about model selection and averaging can also be found in [17, 21, 31–33]. The model averaging method has also been used in many areas of applications, e.g., [4, 5] for forecasting stock market data, [25] for risk of using false models in portfolio

management, [23] for analysis of the Hong Kong housing market, and [26] for a study of phylogenetics in biology. Our development in this article extends the existing theoretical frequentist development to a general framework so it can incorporate biased models under a general setting. Model averaging has been also discussed in the Bayesian framework (see, e.g., [13, 27]). In a Bayesian approach, a weighted average of the posterior distributions under every available candidate model was used for estimation and prediction purposes. The weights were determined by posterior model probabilities. Model averaging in a frequentist setup, as in [11] and also ours, precludes the need to specify any prior distributions, thus removing any possible oversight due to faulty choice of priors. The question in a frequentist setting is how to obtain the weights by a data-driven approach.

The rest of the article is organized as follows. In Section 2, we propose a general framework that covers the framework of [11] as a special case and study asymptotic properties of model averaging estimators. We also derive a consistent estimator for the mean squared error of the model averaging estimator and use it to facilitate our choice of data-driven weights in Subsection 2.4. The development is illustrated in generalized linear models and particularly in linear and logistic model setups. In Section 4, simulation studies are carried out to examine the performance of the proposed estimator and to compare its performance with existing methods.

2 General framework

2.1 Notation and setup

Consider n independent data points $\boldsymbol{y} = (y_1, \ldots, y_n)$ sampled from a distribution having density of the form $f(y) \equiv f(y, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the unknown parameter of interest. Here, the parameter $\boldsymbol{\beta}$ can be written as $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, $p \geqslant 0$, are the parameters that are certainly included in every candidate model and $\boldsymbol{\gamma} \in \mathbb{R}^q$ is the remaining set of parameters that may or may not be included in the candidate models. We assume that p and q are given. As a model averaging method, instead of choosing one particular candidate model as the "correct" model, we consider a set of candidate models, say $\mathcal{M}$, in which each candidate model contains the common parameters $\boldsymbol{\theta}$ and a unique $\boldsymbol{\gamma}'$ that includes m of q components of the parameter $\boldsymbol{\gamma}$, $0 \leqslant m \leqslant q$.

The choice of $\mathcal{M}$ can vary depending on the problem that one is trying to solve. For example, the candidate model set $\mathcal{M}$ can contain all possible 2^q combinations of $\boldsymbol{\gamma}$. Or, one can choose a subset of the 2^q possible models as $\mathcal{M}$. In [9], a set of nested models has been used as candidate models, with $|\mathcal{M}| = q + 1$. In [11], $\mathcal{M}$ includes candidate models that are within an $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. Our development encompasses both setups as there are no restrictions on $\mathcal{M}$, and $\mathcal{M}$ can include any number of candidate models between 1 and 2^q. Similar setup was used in [20] where $\mathcal{M}$ is also unrestricted, but the development there was done in the standard linear regression framework.

Let the parameters in the true model be given by $\boldsymbol{\beta}_{\mathrm{true}} = (\boldsymbol{\theta}_{\mathrm{true}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathrm{true}}^{\mathrm{T}})^{\mathrm{T}}$. Let m^{true} be the number of components of $\boldsymbol{\gamma}$ that are present in the true model. Define $\mathcal{M}_{\in}$ as the collection of the candidate models that contain the true model, thus every model in $\mathcal{M}_{\in}$ contains each and every one of the m^{true} components of $\boldsymbol{\gamma}$. Define $\mathcal{M}_{\notin} = \mathcal{M} - \mathcal{M}_{\in} \subset \mathcal{M}$, so $\mathcal{M}_{\notin}$ contains candidate models for which at least one of those m^{true} components are not present. Clearly, $\mathcal{M} = \mathcal{M}_{\in} \cup \mathcal{M}_{\notin}$.

In [11], a common parameter is also present in all the candidate models. But the treatment of $\boldsymbol{\gamma}$ is different. In particular, the model containing just $\boldsymbol{\theta}$ is called a "narrow model" and the true model is chosen of the form $f(\boldsymbol{y}) = f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}_0 + \delta/\sqrt{n})$. Here, parameter δ determines how far a candidate model can vary from the narrow model and $\boldsymbol{\gamma}_0$ is a given value of $\boldsymbol{\gamma}$ for which any extended model reduces down to the narrow model. Thus, this choice of true model essentially requires that the all candidate models are within an $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. Any model that is beyond $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model is excluded from the analysis. In this paper, we remove this rather restrictive constraint. Indeed, we assume the parameter for the true model is $\boldsymbol{\beta}_{\mathrm{true}} = (\boldsymbol{\theta}_{\mathrm{true}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathrm{true}}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\gamma}_{\mathrm{true}}$ may or may not have any of the q components, and the candidate model set $\mathcal{M}$ can be a subset or contain all possible 2^q combinations of $\boldsymbol{\gamma}$. Thus in our model setup there are no restrictions on the choice of true model or

on the set of candidate models as in [11]. Furthermore, we can treat the setup considered in [11] as a special case of ours by restricting $\boldsymbol{\gamma}_{\text{true}}$ so that all the candidate models have a bias of order $\mathcal{O}(1/\sqrt{n})$ or less.

Note that every candidate model includes a unique $\boldsymbol{\gamma}$ that may or may not include all q components. Thus the numbers of parameters from different candidate models may be different. For ease of presentation and following [9], we introduce an augmentation scheme to bring all of them to the same length. We first illustrate the idea using the regression example considered by Hansen [9]: $\boldsymbol{y}$ is the vector of responses, $\boldsymbol{X}$ is the design matrix with full column rank $p+q$ and the candidate models are nested models. We further assume the first p columns of $\boldsymbol{X}$ are always included in the candidate models; the special case with $p=0$ goes back to the setup of [9]. It follows that the k-th candidate model includes the first $p+k$ columns of $\boldsymbol{X}$, $k=0,\ldots,q$. Denote by $\widehat{\boldsymbol{\beta}}_k$ the estimated regression parameters corresponding to the k-th candidate model. Then the $(p+k)\times 1$ vector $\widehat{\boldsymbol{\beta}}_k$ can be augmented to a $(p+q)\times 1$ vector $(\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}$, by adding $(q-k)$ 0's. The augmented estimator for the k-th candidate model is given by

$$\widetilde{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}} = \begin{bmatrix} (\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{y} \\ 0 \end{bmatrix} \tag{2.1}$$

(see, e.g., [9] adopted this augmentation on a set of nested candidate models).

More generally, let $\boldsymbol{\beta}_k$ be the parameter for the k-th model in $\mathcal{M}$. Assume the length of $\boldsymbol{\beta}_k$ is $p+m_k$, where m_k depends on k. Define the log-likelihood for the i-th observation in the k-th model as

$$\ell_{k;i}(\boldsymbol{\beta}_k) = \log f(y_i, \boldsymbol{\beta}_k).$$

The maximum likelihood estimate (MLE) of $\boldsymbol{\beta}_k$ with the k-th model is $\widehat{\boldsymbol{\beta}}_k = \arg\max_{\beta_k} \ell_k(\boldsymbol{\beta}_k)$, where $\ell_k(\boldsymbol{\beta}_k) = \sum_{i=1}^n \ell_{k;i}(\boldsymbol{\beta}_k)$. Write the score function of the k-th model as $S_k(\boldsymbol{\beta})$. As in the example above, the vector $\boldsymbol{\beta}_k$ for the k-th model can be augmented to a $(p+q)\times 1$ vector $(\boldsymbol{\beta}_k^{\mathrm{T}}, \boldsymbol{c}_k^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{c}_k$ is a fixed value used for augmentation to hold spaces. The augmented maximum likelihood estimator is given by $\widetilde{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}, \boldsymbol{c}_k^{\mathrm{T}})^{\mathrm{T}}$. The fixed value augmentation does not affect the parameter, and only appends the length of the parameter. In the linear model example above the values $\boldsymbol{c}_k = 0$. Some examples of $\boldsymbol{c}_k \neq 0$ can be found in [24]. Similarly, $\boldsymbol{\beta}_{\text{true}}$ can be augmented to a $(p+q)\times 1$ vector $(\boldsymbol{\beta}_{\text{true}}^{\mathrm{T}}, \boldsymbol{c}^{\mathrm{T}})^{\mathrm{T}}$ for a certain fixed set of $\boldsymbol{c}$ without altering the true model. Thus, without loss of generality and from now on, we assume $\boldsymbol{\beta}_{\text{true}}$ is a $(p+q)\times 1$ vector in the sense that some of the elements may be augmented to fill the space.

For the model $k\in\mathcal{M}$, let us define $\boldsymbol{\beta}_k^* \in \mathbb{R}^{p+m_k}$ as the solution of the equation $\mathrm{E}S_k(\boldsymbol{\beta}) = 0$, where $S_k(\boldsymbol{\beta})$ is the score function of the k-th model having $p+m_k$ parameters. Define, as before, $\widetilde{\boldsymbol{\beta}}_k^* \in \mathbb{R}^{p+q}$ as the $\boldsymbol{c}$-augmented version of $\boldsymbol{\beta}_k^*$. Since the score function is Fisher consistent, $\widetilde{\boldsymbol{\beta}}_k \to \widetilde{\boldsymbol{\beta}}_k^*$ under usual regularity conditions. But this $\widetilde{\boldsymbol{\beta}}_k^*$ may not be close to $\boldsymbol{\beta}_{\text{true}}$.

Let $\boldsymbol{\mu}: \mathbb{R}^{p+q} \to \mathbb{R}^\ell$ be a general function that is 1st order partially differentiable and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta}_{\text{true}})$ is the parameter of interest. Then, the model averaging estimator of $\boldsymbol{\mu}$ is defined as

$$\widehat{\boldsymbol{\mu}}_{\text{ave}} = \sum_{k\in\mathcal{M}} w_k \boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k), \tag{2.2}$$

where the weights $0 \leqslant w_k \leqslant 1, \forall\, k$, and $\sum_{k\in\mathcal{M}} w_k = 1$. In the remainder of this section, we derive the asymptotic properties of the model averaging estimator (2.2) for any given set of weights w_k.

2.2 Main results

We assume the usual regularity conditions under which the familiar likelihood asymptotic arguments apply (see the conditions listed in Appendix A). See also [18, 19, 30] for more details.

Let $\nabla\boldsymbol{\mu}\in\mathbb{R}^{\ell\times(p+q)}$ be the first order derivative of the $\mathbb{R}^{p+q}\to\mathbb{R}^{\ell}$ function $\boldsymbol{\mu}$. Define

$$\boldsymbol{H}_k=\lim_{n\to\infty}\frac{1}{n}\mathrm{E}[\ell_k''(\boldsymbol{\beta}_k^*)],$$

and assume it is invertible. We also assume

$$\text{(A1)}\quad \lim_n\frac{1}{n}\sum_{i=1}^n\mathrm{E}\Big[\max_{k\in\mathcal{M}}\|\nabla\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)\boldsymbol{H}_k^{-1}\ell_{k;i}'(\boldsymbol{\beta}_k^*)\|^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}\|\nabla\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)\boldsymbol{H}_k^{-1}\ell_{k;i}'(\boldsymbol{\beta}_k^*)\|>\sqrt{n}\epsilon\Big\}\Big]=0,$$

for any $\epsilon>0$, where $\mathbb{I}\{\cdot\}$ is the indicator function. We have the following theorem. Its proof is given in Appendix A.

Theorem 2.1. *Let $\widetilde{\boldsymbol{\beta}}_k$ be the $\boldsymbol{c}$-augmented MLE as defined in* (2.1) *for the k-th model in $\mathcal{M}$. Let $0\leqslant w_k\leqslant 1$ for $k\in\mathcal{M}$ be model weights such that $\sum_k w_k=1$. Assume Condition* (A1) *holds. Then, the asymptotic distribution of the model averaging estimator for $\boldsymbol{\mu}(\boldsymbol{\beta}_{\mathrm{true}})$ is given as,*

$$\sqrt{n}\sum_{k\in\mathcal{M}}w_k\{\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k)-\boldsymbol{\mu}(\boldsymbol{\beta}_{\mathrm{true}})\}-\sqrt{n}\sum_{k\in\mathcal{M}}w_k\{\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)-\boldsymbol{\mu}(\boldsymbol{\beta}_{\mathrm{true}})\}\xrightarrow{\mathrm{D}}\mathsf{N}(0,\boldsymbol{\Sigma}_w),\tag{2.3}$$

where the variance $\boldsymbol{\Sigma}_w$ is given by

$$\boldsymbol{\Sigma}_w=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathrm{E}\Big[\Big\{\sum_k w_k\nabla\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)^{\mathrm{T}}\boldsymbol{H}_k^{-1}\ell_{k;i}'\Big\}^{\otimes 2}\Big].\tag{2.4}$$

Condition (A1) implies that the contribution of $\nabla\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)\boldsymbol{H}_k^{-1}\ell_{k;i}'(\boldsymbol{\beta}_k^*)$ to the total variance, for each model k in the set $\mathcal{M}$ and for each $1\leqslant i\leqslant n$ is asymptotically negligible, and it is satisfied in a wide array of cases. We provide a set of sufficient conditions under which it is satisfied, and we also provide such examples in the cases of linear and logistic models in Subsection 2.4. See further discussions of the condition in Subsection 2.4.

In our general framework, there is no guarantee that $\widetilde{\boldsymbol{\beta}}_k^*=\boldsymbol{\beta}_{\mathrm{true}}$, neither does $\|\widetilde{\boldsymbol{\beta}}_k^*-\boldsymbol{\beta}_{\mathrm{true}}\|\to 0$ asymptotically, particularly when $k\in\mathcal{M}_{\notin}$. So $\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*)-\boldsymbol{\mu}(\boldsymbol{\beta}_{\mathrm{true}})$ is not necessarily 0, even asymptotically. But we can view it as a measurement of the bias by the k-th model. Thus, with the second term on the left-hand side of (2.3) serving as a bias correction term, Theorem 2.1 states that the model averaging estimator still retains the usual form of asymptotic normality after the bias correction.

Note that, in the theorem, the weights are fixed and non-random, which is different from [11] which also considered random weights. In practice (see Subsection 2.4), we often estimate the weights using data. Strictly speaking, the result in the theorem does not directly apply to our proposed model averaging estimator. Nevertheless, it is still useful in motivating the practical estimator by choosing weights that minimize the asymptotic variance.

All the candidate models have $\boldsymbol{\theta}$ in common. We can use Theorem 2.1 to construct asymptotic convergence results for the common parameter $\boldsymbol{\theta}$. If we consider a function from $(\boldsymbol{\theta}^{\mathrm{T}},\boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}\mapsto\boldsymbol{\theta}$ to extract the $\boldsymbol{\theta}$ parameter, then by a direct application of Theorem 2.1 we can derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ as given below in Corollary 2.2.

Corollary 2.2. *Let $\boldsymbol{\theta}$ be the common parameter for all candidate models in $\mathcal{M}$. Let $\boldsymbol{\beta}_k^*=(\boldsymbol{\theta}_k^{*\mathrm{T}},\boldsymbol{\gamma}_k^{*\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}_k=(\widehat{\boldsymbol{\theta}}_k^{\mathrm{T}},\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}})^{\mathrm{T}}$. Then under the same setup as in Theorem* 2.1*,*

$$\sqrt{n}\sum_{k\in\mathcal{M}}w_k(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_{\mathrm{true}})-\sqrt{n}\sum_{k\in\mathcal{M}}w_k(\boldsymbol{\theta}_k^*-\boldsymbol{\theta}_{\mathrm{true}})\xrightarrow{\mathrm{D}}\mathsf{N}(0,\boldsymbol{\Sigma}_w),\tag{2.5}$$

where the variance is given by

$$\boldsymbol{\Sigma}_w=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathrm{E}\Big\{\Big(\sum_k w_k[\boldsymbol{I}_p,\boldsymbol{0}]\boldsymbol{H}_k^{-1}\ell_{k;i}'\Big)^{\otimes 2}\Big\}.$$

2.3 Connection to the development in [11]: Hjort and Claeskens (2003)

The development of [11] required that all candidate models are within an $\mathcal{O}(1/\sqrt{n})$ neighborhood of the true model. We broaden this framework in our development. In particular, we show in this subsection that the results described in [11] can be obtained as a special case of our result. Again, we would like to point out that the latter theoretically allows random weights while in the paper we require the weights to be non-random. Thus, technically we only recover the result in [11] when the weights are deterministic.

We start with a description of the mis-specified model setup used in [11]. Let $Y_1, \ldots, Y_n$ be an independent and identically distributed (i.i.d.) sample from density f of maximum $p+q$ parameters. The parameter of interest is $\mu = \mu(f)$, where $\mu : \mathbb{R}^{p+q} \to \mathbb{R}$. The model that includes just p parameters, say $\boldsymbol{\theta}$, is defined as the narrow model, while any extended model $f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ reduces to the narrow model for $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$; here the vector $\boldsymbol{\gamma}_0$ is fixed and known. For the k-th model with unknown parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma}_k)$, the MLE of μ is written as $\widehat{\mu}_k = \mu(\widehat{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\gamma}}_k, \boldsymbol{\gamma}_{0,k^c})$, where k^c refers to the elements that are not contained in $\boldsymbol{\gamma}_k$. Thus in this setup, if a parameter γ_j is not included in the candidate model, we set $\gamma_j = \gamma_{j,0}$, the j-th element of $\boldsymbol{\gamma}_0$. The true model is assumed to be

$$f_{\text{true}}(y) = f(y, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0 + \boldsymbol{\delta}/\sqrt{n}), \tag{2.6}$$

where $\boldsymbol{\delta}$ signifies the deviation of the model in directions $1, \ldots, q$. So $\boldsymbol{\beta}_{\text{true}} = (\boldsymbol{\theta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}} + \boldsymbol{\delta}^{\mathrm{T}}/\sqrt{n})^{\mathrm{T}}$. Let us write $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}$. We will also write $\mu_{\text{true}} = \mu(\boldsymbol{\beta}_{\text{true}})$, which is the estimand of interest. Under this model setup, Hjort and Claeskens [11] derived asymptotic normality result for the model averaging estimator $\sum_k w_k \widehat{\mu}_k$. To describe their result, let us first define

$$S(y) = \begin{bmatrix} U(y) \\ V(y) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ \frac{\partial}{\partial \boldsymbol{\gamma}} \log f(y, \boldsymbol{\theta}, \boldsymbol{\gamma}) \end{bmatrix} \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \boldsymbol{\gamma}=\boldsymbol{\gamma}_0} \quad \text{and} \quad \text{var}\{S(Y)\} = \begin{bmatrix} \boldsymbol{J}_{00} & \boldsymbol{J}_{01} \\ \boldsymbol{J}_{01} & \boldsymbol{J}_{11} \end{bmatrix} = \boldsymbol{J}_{\text{full}}.$$

Let $\overline{U}_n = n^{-1} \sum_i U(Y_i)$ and $\overline{V}_n = n^{-1} \sum_i V(Y_i)$. Denote by $V_k(Y)$ and $\overline{V}_{n;k}$ the appropriately subsetted vectors obtained from $V(Y)$ and $\overline{V}_n$, with the subset indices corresponding to that of $\widehat{\boldsymbol{\gamma}}$ in model $k \in \mathcal{M}$, respectively. Also, define $\boldsymbol{J}_k = \text{var}\{U(Y), V_k(Y)\}$ for all $k \in \mathcal{M}$. Hjort and Claeskens [11] showed that,

$$\sqrt{n}\Big(\sum_k w_k \widehat{\mu}_k - \mu_{\text{true}} \Big) \xrightarrow{\text{D}} \sum_k w_k \Lambda_k, \tag{2.7}$$

where

$$\Lambda_k = \begin{pmatrix} \partial \mu(\boldsymbol{\beta}_0)/\partial \boldsymbol{\theta} \\ \partial \mu(\boldsymbol{\beta}_0)/\partial \boldsymbol{\gamma}_k \end{pmatrix}^{\mathrm{T}} \left\{ \boldsymbol{J}_k^{-1} \begin{pmatrix} \boldsymbol{J}_{01} \boldsymbol{\delta} \\ \pi_k \boldsymbol{J}_{11} \boldsymbol{\delta} \end{pmatrix} + \boldsymbol{J}_k^{-1} \begin{pmatrix} M \\ N_k \end{pmatrix} \right\} - \left(\frac{\partial \mu(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\gamma}} \right)^{\mathrm{T}} \boldsymbol{\delta},$$

where $(M, N_k) \sim \mathsf{N}_{p+q}(\boldsymbol{0}, \boldsymbol{J}_k)$. Here, $\pi_k \in \mathbb{R}^{|M_k| \times q}$ is the projection matrix that projects any vector $\boldsymbol{u} \in \mathbb{R}^q$ to $\boldsymbol{u}_k \in \mathbb{R}^{|M_k|}$ with indices as given by $M_k \in \mathcal{M}$.

The following corollary states that the result in (2.7) can be directly obtained from Theorem 2.1 and thus Theorem 2.1 covers the special setting (2.6) of [11]. A proof of the corollary can be found in Appendix A.

Corollary 2.3. *Under the mis-specification model* (2.6)*, the asymptotic bias and variance in* (2.7) *matches those in Theorem* 2.1*.*

2.4 Selection of weights in frequentist model averaging

Model averaging acknowledges the uncertainty caused by model selection and tackles the problem by weighting all models under consideration. To make it effective, it is desirable that the weights can reflect the impact of each candidate model, which can be achieved by properly assigning a weight to each candidate model. If model k' is more likely to impact or is more plausible than the model k, its associated

weight $w_{k'}$ should be no smaller than w_k for the model k. In our development, we propose to measure the strength of a model by its mean squared error, based on which we obtain a set of data-adaptive weights by minimizing the mean squared error of the combined model averaging estimator. A similar scheme was developed in [20], where the authors minimized an unbiased estimator of mean squared error to obtain their optimal weights. As in [20], we assume that the true model is included in the set of candidate models in the development of our weighing scheme.

Recall Theorem 2.1, and the asymptotic mean squared error (AMSE) of $\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k)$ is,

$$Q(\boldsymbol{w}) = \operatorname{trace}\bigg(\bigg[\sum_{k\in\mathcal{M}} w_k\{\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k^*) - \boldsymbol{\mu}(\boldsymbol{\beta}_{\text{true}})\}\bigg]^{\otimes 2} + \frac{1}{n}\boldsymbol{\Sigma}_w\bigg), \tag{2.8}$$

for any given set of weights. However, this quantity depends on some unknown parameters, so we instead consider its estimate

$$\widehat{Q}_n(\boldsymbol{w}) = Q(\boldsymbol{w})\,|_{\boldsymbol{\beta}_{\text{true}}=\widehat{\boldsymbol{\beta}}_{\text{cons}},\, \tilde{\boldsymbol{\beta}}_k^*=\hat{\tilde{\boldsymbol{\beta}}}_k^*,\, \boldsymbol{\Sigma}_w=\hat{\boldsymbol{\Sigma}}_w}.$$

Here, we assume that we can consistently estimate $\boldsymbol{\beta}_{\text{true}}$, $\tilde{\boldsymbol{\beta}}_k^*$ and $\boldsymbol{\Sigma}_w$, say, by $\widehat{\boldsymbol{\beta}}_{\text{cons}}$, $\hat{\tilde{\boldsymbol{\beta}}}_k^*$, $\hat{\boldsymbol{\Sigma}}_w$, respectively. Note that, to compute $\widehat{Q}_n(\boldsymbol{w})$, we only need a consistent estimate for $\boldsymbol{\beta}_{\text{true}}$ and $\widehat{\boldsymbol{\beta}}_{\text{cons}}$ does not need to be efficient. So we often obtain it using the full model. For consistently estimating $\tilde{\boldsymbol{\beta}}_k^*$ and $\boldsymbol{\Sigma}_w$ in linear and logistic models, we defer it to Subsections 3.1 and 3.2 with details. We propose to obtain a set of data adaptive weights $\boldsymbol{w}_n^*$ by minimizing $\widehat{Q}_n(\boldsymbol{w})$:

$$\boldsymbol{w}_n^* = \arg\min_{\boldsymbol{w}} \widehat{Q}_n(\boldsymbol{w}).$$

The numerical performance of the proposed averaging estimators will be evaluated in Section 4. In the next section, we illustrate the procedure in the linear and logistic models in details.

3 Model averaging and weight selection in regression models

We now discuss the model averaging estimator described in Section 2 for generalized linear models (GLM). Specifically, let $\mathrm{E}y_i = g(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, where g is a given link function connecting the mean and the linear predictor $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. We consider a set $\mathcal{M}$ of 2^q models. Suppose we want to estimate a function $\boldsymbol{\mu}(\boldsymbol{\beta})$ and, as defined in (2.2), the final model averaging estimator is given by $\widehat{\boldsymbol{\mu}}_{\text{ave}} = \sum_{k\in\mathcal{M}} w_k\boldsymbol{\mu}(\widetilde{\boldsymbol{\beta}}_k)$. Since the set up for Theorem 2.1 is for a general parametric model, the same asymptotic convergence results hold for GLM. In particular we verify Condition (A1) and discuss the data-driven weight choices below in two special cases: linear and logistic regression models.

3.1 Prediction in linear regression models

We first derive the model averaging estimator in the linear regression framework:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ is a non-random design matrix of full column rank; i.e., $\operatorname{rank}(\boldsymbol{X}) = p+1$, and $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$.

Let $\mathcal{M} = \{M_k\}_{k=1}^{|\mathcal{M}|}$ be the set of candidate models. Here, M_k denotes a particular set of features having cardinality $|M_k|$. Define $\boldsymbol{X}_k \in \mathbb{R}^{n\times|M_k|}, 1 \leqslant k \leqslant |\mathcal{M}|$ as the design matrix of the k-th candidate model with the features in M_k. We consider zero-augmentation of the parameter set $\boldsymbol{\beta}_k$ for all k. Let $\widetilde{\boldsymbol{X}}_k \in \mathbb{R}^{n\times(p+1)}$ be the augmented version of $\boldsymbol{X}_k$ with the missing columns replaced by the $\boldsymbol{0}$ vector. In our analysis, all the candidate models contain the intercept term corresponding to β_0. With the rest of the p components, we can construct 2^p candidate models, all of which are included in our analysis.

Let us fix an $\boldsymbol{x}^* \in \mathbb{R}^{p+1}$. Define $\boldsymbol{x}_k^* \in \mathbb{R}^{|M_k|}$ so that $\boldsymbol{x}_k^*$ consists of those components of $\boldsymbol{x}^*$ indexed by $M_k \in \mathcal{M}$. Consider the particular choice of the function $\mu : \mathbb{R}^{p+1} \to \mathbb{R}$ so that for $\boldsymbol{b} \in \mathbb{R}^{p+1}$, $\mu(\boldsymbol{b}) = \boldsymbol{x}^{*\mathrm{T}}\boldsymbol{b}$.

Clearly, we have $\nabla\mu(\boldsymbol{\beta}) = \boldsymbol{x}^*$. For the following discussion, we are interested in the model averaging estimator of $\mu(\boldsymbol{\beta}_{\text{true}}) = \boldsymbol{x}^{*\text{T}}\boldsymbol{\beta}_{\text{true}}$, which is given by $\widehat{\mu}_{\text{ave}} = \sum_k w_k \boldsymbol{x}_k^{*\text{T}}\widehat{\boldsymbol{\beta}}_k$ with $w_k \geqslant 0$ and $\sum_k w_k = 1$. In the simulations, we will use $\boldsymbol{x}^*$ generated from the known covariate distribution, while for the real data, we split the whole data set into a training set and a test set and $\boldsymbol{x}^*$ will be set to be the covariate in the test set.

For the k-th candidate model with $\boldsymbol{\beta}_k \in \mathbb{R}^{|M_k|}$, the score function is given by $\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k) = \boldsymbol{X}_k^{\text{T}}(\boldsymbol{y} - \boldsymbol{X}_k\boldsymbol{\beta}_k)$ and $\boldsymbol{H}_k$ is given by $\boldsymbol{H}_k = -(1/n)\boldsymbol{X}_k^{\text{T}}\boldsymbol{X}_k$; note that this follows from the definition immediately preceding Condition (A1). Thus our Hessian matrix satisfies the condition as it does not depend on $\boldsymbol{y}$. Similarly we note that regarding Condition (A1),

$$|\nabla\mu(\widetilde{\boldsymbol{\beta}}_k^*)\boldsymbol{H}_k^{-1}\ell_{k;i}'(\boldsymbol{\beta}_k^*)| = |(y_i - [\boldsymbol{X}_k]_{i,\bullet}^{\text{T}}\boldsymbol{\beta}_k^*)\ \boldsymbol{x}_k^{*\text{T}}(\boldsymbol{X}_k^{\text{T}}\boldsymbol{X}_k/n)^{-1}[\boldsymbol{X}_k]_{i,\bullet}| = |c_{ik}(\varepsilon_i + A_{ik})|,$$

where $c_{ik} = \boldsymbol{x}_k^{*\text{T}}(\boldsymbol{X}_k^{\text{T}}\boldsymbol{X}_k/n)^{-1}[\boldsymbol{X}_k]_{i,\bullet}$ and $A_{ik} = \boldsymbol{x}_i^{\text{T}}\boldsymbol{\beta}_{\text{true}} - [\boldsymbol{X}_k]_{i,\bullet}^{\text{T}}\boldsymbol{\beta}_k^*$ are fixed constants, and $[\boldsymbol{X}_k]_{i,\bullet}$ is the i-th column of the matrix $\boldsymbol{X}_k^{\text{T}}$.

Note that $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$. Then the key term in Condition (A1) can now be written, for any arbitrary $\epsilon > 0$, as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big\{\max_{k\in\mathcal{M}}|c_{ik}(\varepsilon_i + A_{ik})|\Big\}^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}|c_{ik}(\varepsilon_i + A_{ik})| > \sqrt{n}\epsilon\Big\}\\
&\quad\leqslant \max_{1\leqslant i\leqslant n}\mathrm{E}\Big\{\max_{k\in\mathcal{M}}|c_{ik}(\varepsilon_i + A_{ik})|\Big\}^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}|c_{ik}(\varepsilon_i + A_{ik})| > \sqrt{n}\epsilon\Big\}\\
&\quad\leqslant \max_{1\leqslant i\leqslant n}\Big\{\max_{k}|c_{ik}|^2\Big\}\mathrm{E}\Big\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}|\Big\}^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}| > \sqrt{n}\Big(\epsilon\Big/\max_{k}|c_{ik}|\Big)\Big\}.
\end{aligned}$$

Moreover, if $|c_{ik}| \leqslant C$ for some fixed constant $C > 0$, then it would further suffice to prove that,

$$\lim_{n\to\infty}\max_{1\leqslant i\leqslant n}\mathrm{E}\Big\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}|\Big\}^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon'\Big\} = 0,$$

where $\epsilon' = \epsilon/C$ is a fixed constant. It is appropriate to note that we can have a bound of c_{ik} as

$$\max_k|c_{ik}| = \max_k|\boldsymbol{x}_k^{*\text{T}}(\boldsymbol{X}_k^{\text{T}}\boldsymbol{X}_k/n)^{-1}[\boldsymbol{X}_k]_{i,\bullet}| \leqslant \|\boldsymbol{x}^*\|\|\boldsymbol{x}_i\|\max_k\frac{1}{\lambda_{\min}(\boldsymbol{X}_k^{\text{T}}\boldsymbol{X}_k/n)}.$$

Here, $\lambda_{\min}(\boldsymbol{B})$ denotes the smallest singular value of matrix $\boldsymbol{B}$. Now by an application of the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathrm{E}\Big\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}|\Big\}^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon'\Big\}\\
&\quad\leqslant \Big\{\mathrm{E}\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}|^4\Big\}^{1/2}\Big\{\mathrm{P}\Big(\max_{k\in\mathcal{M}}|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon'\Big)\Big\}^{1/2}\\
&\quad\leqslant \Big\{\sum_{k\in\mathcal{M}}\mathrm{E}(\varepsilon_i + A_{ik})^4\Big\}^{1/2}\Big\{\sum_{k\in\mathcal{M}}\mathrm{P}(|\varepsilon_i + A_{ik}| > \sqrt{n}\epsilon')\Big\}^{1/2}\\
&\quad\leqslant \Big\{\sum_{k\in\mathcal{M}}(A_{ik}^4 + 6A_{ik}^2\sigma^2 + 3\sigma^4)\Big\}^{1/2}\Big\{\sum_{k\in\mathcal{M}}\mathrm{P}(|\varepsilon_i| > \sqrt{n}\epsilon' - |A_{ik}|)\Big\}^{1/2}. \qquad (3.1)
\end{aligned}$$

Thus it follows that for $|\mathcal{M}|$ finite, as n goes to infinity, the right-hand side of (3.1) goes to zero and thus Condition (A1) is satisfied.

The MLE of β_k in the k-th model is given by

$$\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}_k^{\text{T}}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^{\text{T}}\boldsymbol{y}.$$

Let $\boldsymbol{\beta}_k^*$ be such that $\mathrm{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k^*)=\mathbf{0}$; $\mathrm{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k)$ being the score function of the k-th model, solving which we find that,

$$\boldsymbol{\beta}_k^*=(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{true}}. \tag{3.2}$$

As discussed in Subsection 2.1, the entire set of candidate models can be divided into two categories. The 1st category contains the ones that are biased and is denoted by $\mathcal{M}_{\notin}$ and the second category contains ones that are not and is denoted by $\mathcal{M}_{\in}$. So, for $k\in\mathcal{M}_{\in}$ we have $\boldsymbol{\beta}_k^*=\boldsymbol{\beta}_{\mathrm{true}}$, whereas for $k\in\mathcal{M}_{\notin}$ we have $\boldsymbol{\beta}_k^*\neq\boldsymbol{\beta}_{\mathrm{true}}$. Therefore the bias term of model averaging estimator $\widehat{\mu}_{\mathrm{ave}}$ can be written as,

$$\sum_{k\in\mathcal{M}_{\notin}}w_k(\boldsymbol{x}_k^{*\mathrm{T}}\boldsymbol{\beta}_k^*-\boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}_{\mathrm{true}})=\sum_{k\in\mathcal{M}_{\notin}}w_k\boldsymbol{x}_k^{*\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{true}}-\boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}_{\mathrm{true}}.$$

Since the weights assigned to the models are unknown, we propose an estimate of the mean squared error (MSE) and minimize the MSE to obtain weights that would be assigned to the candidate models. From Theorem 2.1, the asymptotic mean squared error (AMSE) of $\widehat{\mu}_{\mathrm{ave}}$ is given by

$$\begin{aligned}Q(\boldsymbol{w})=\Bigg[&\Bigg\{\sum_{k\in\mathcal{M}_{\notin}}w_k(\boldsymbol{x}_k^{*\mathrm{T}}\boldsymbol{\beta}_k^*-\boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}_{\mathrm{true}})\Bigg\}^2\\&+\frac{1}{n^2}\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}}w_kw_{k'}\boldsymbol{x}_k^{*\mathrm{T}}\boldsymbol{H}_k^{-1}\mathrm{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_{\mathrm{true}})\boldsymbol{\ell}_{k'}'(\boldsymbol{\beta}_{\mathrm{true}})^{\mathrm{T}}\boldsymbol{H}_{k'}^{-1^{\mathrm{T}}}\boldsymbol{x}^*{}_{k'}\Bigg].\end{aligned}$$

Since $\boldsymbol{H}_k$ does not depend on $\boldsymbol{y}$, we focus on $\mathrm{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_{\mathrm{true}})\boldsymbol{\ell}_{k'}'(\boldsymbol{\beta}_{\mathrm{true}})^{\mathrm{T}}$, which equals

$$\boldsymbol{X}_k^{\mathrm{T}}\mathrm{E}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{true}})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{true}})^{\mathrm{T}}\boldsymbol{X}_{k'}=\sigma^2\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_{k'}.$$

It follows that

$$\begin{aligned}Q(\boldsymbol{w})=\Bigg\{&\sum_{k\in\mathcal{M}_{\notin}}\sum_{k'\in\mathcal{M}_{\notin}}w_kw_{k'}(\boldsymbol{x}_k^{*\mathrm{T}}\boldsymbol{\beta}_k^*-\boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}_{\mathrm{true}})(\boldsymbol{x}^{*\mathrm{T}}_{k'}\boldsymbol{\beta}_{k'}^*-\boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}_{\mathrm{true}})\\&+\sigma^2\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}}w_kw_{k'}\boldsymbol{x}_k^{*\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)^{-1}\ \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_{k'}\ (\boldsymbol{X}_{k'}^{\mathrm{T}}\boldsymbol{X}_{k'})^{-1}\boldsymbol{x}^*{}_{k'}\Bigg\}.\end{aligned}$$

Define the estimates of $\boldsymbol{\beta}$ and σ as $\widehat{\boldsymbol{\beta}}_{\mathrm{full}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ and $\widehat{\sigma}^2_{\mathrm{full}}=\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{full}}\|^2/n$, respectively. Then $(\widehat{\boldsymbol{\beta}}_{\mathrm{full}},\widehat{\sigma}_{\mathrm{full}})$ are consistent estimates of $(\boldsymbol{\beta}_{\mathrm{true}},\sigma)$ under mild conditions. We therefore propose to estimate $Q(\boldsymbol{w})$ by

$$\begin{aligned}\widehat{Q}(\boldsymbol{w})=&\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}}w_kw_{k'}(\boldsymbol{x}_k^{*\mathrm{T}}\widehat{\boldsymbol{\beta}}_k-\boldsymbol{x}^{*\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{full}})(\boldsymbol{x}^{*\mathrm{T}}_{k'}\widehat{\boldsymbol{\beta}}_{k'}-\boldsymbol{x}^{*\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{full}})\\&+\widehat{\sigma}^2_{\mathrm{full}}\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}}w_kw_{k'}\boldsymbol{x}_k^{*\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)^{-1}\ \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_{k'}(\boldsymbol{X}_{k'}^{\mathrm{T}}\boldsymbol{X}_{k'})^{-1}\boldsymbol{x}_k^*,\end{aligned} \tag{3.3}$$

where the second term above is an estimator of $(1/n)\Sigma_w$. We obtain the weights for model averaging estimator $\boldsymbol{w}=(w_1,\ldots,w_{|\mathcal{M}|})$ such that $\widehat{Q}(\boldsymbol{w})$ in (3.3) is minimized.

3.2 Estimation in logistic regression framework

In this section we study the proposed model averaging estimation method under logistic regression models. Let $\boldsymbol{y}\in\mathbb{R}^n$ be n independent copies of a dichotomous response variable Y taking values 0/1. Let $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\mathrm{T}}\in\mathbb{R}^{n\times(p+1)}$ be a set of features. The logit model is given by,

$$p_i=\mathrm{P}(y_i=1\,|\,\boldsymbol{X})=\frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})},\quad\forall\,i=1,\ldots,n,$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is the vector of unknown parameters of interest. The log-likelihood for the logistic regression can be written as,

$$\ell_k(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}) = \log \prod_{i=1}^{n} \frac{\exp(y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})} = \sum_{i=1}^{n} y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \sum_{i=1}^{n} \log(1 + \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})).$$

As before, let $\mathcal{M} = \{M_k\}_{k=1}^{|\mathcal{M}|}$ be the set of candidate models. Here, M_k denotes a particular set of features having cardinality $|M_k|$. Define $\boldsymbol{X}_k \in \mathbb{R}^{n \times |M_k|}, 1 \leqslant k \leqslant |\mathcal{M}|$ as the design matrix of the k-th candidate model with the features in M_k. Denote by $[\boldsymbol{X}_k]_{i,\bullet}$ the i-th column of the matrix $\boldsymbol{X}_k$, thus $[\boldsymbol{X}_k]_{i,\bullet} \in R^{|M_k|}$. Let $\boldsymbol{\beta}_k \in \mathbb{R}^{|M_k|}$ be the parameter vector with components corresponding to the index set M_k. We consider zero-augmentation of the parameter set $\boldsymbol{\beta}_k$ for all k as was done for the linear regression models.

Again, we consider estimation of a function of the form $p : \mathbb{R}^{p+1} \to \mathbb{R}$ given by

$$p(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}^{*\mathrm{T}} \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^{*\mathrm{T}} \boldsymbol{\beta})}. \tag{3.4}$$

Let the unknown true parameter in our model be $\boldsymbol{\beta}_{\mathrm{true}} \in \mathbb{R}^{p+1}$. Then

$$\boldsymbol{p}_{\mathrm{true}} = \boldsymbol{p}(\boldsymbol{\beta}_{\mathrm{true}}) := \exp(\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{true}})/\{1 + \exp(\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{true}})\} \in \mathbb{R}^n$$

calculated componentwise. To estimate the parameter $p_{\mathrm{true}} = p(\boldsymbol{\beta}_{\mathrm{true}})$, we consider the model averaging estimator given by $\widehat{p}_{\mathrm{ave}} = \sum_{k \in \mathcal{M}} w_k p(\widetilde{\boldsymbol{\beta}}_k)$, where $\widetilde{\boldsymbol{\beta}}_k$ is the 0-augmented version of the MLE $\widehat{\boldsymbol{\beta}}_k$ of $\boldsymbol{\beta}_k$ for the k-th model. The score function for the k-th model is given by

$$\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k) = \sum_i y_i [\boldsymbol{X}_k]_{i,\bullet} - \sum_i \frac{\exp([\boldsymbol{X}_k]_{i,\bullet}^{\mathrm{T}} \boldsymbol{\beta}_k)}{1 + \exp([\boldsymbol{X}_k]_{i,\bullet}^{\mathrm{T}} \boldsymbol{\beta}_k)} [\boldsymbol{X}_k]_{i,\bullet} = \boldsymbol{X}_k^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{p}_k), \quad \forall\, 1 \leqslant k \leqslant |\mathcal{M}|,$$

where $\boldsymbol{p}_k = \exp(\boldsymbol{X}_k \boldsymbol{\beta}_k)/\{1 + \exp(\boldsymbol{X}_k \boldsymbol{\beta}_k)\} \in \mathbb{R}^n$. The second derivative of the log-likelihood is given by

$$\boldsymbol{\ell}_k''(\boldsymbol{\beta}_k) = \sum_{i=1}^{n} \frac{\exp([\boldsymbol{X}_k]_{i,\bullet}^{\mathrm{T}} \boldsymbol{\beta}_k)}{\{1 + \exp([\boldsymbol{X}_k]_{i,\bullet}^{\mathrm{T}} \boldsymbol{\beta}_k)\}^2} [\boldsymbol{X}_k]_{i,\bullet} [\boldsymbol{X}_k]_{i,\bullet}^{\mathrm{T}} = \boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{W}_k (\boldsymbol{I}_n - \boldsymbol{W}_k) \boldsymbol{X}_k, \quad \forall\, 1 \leqslant k \leqslant |\mathcal{M}|,$$

where the weight matrix $\boldsymbol{W}_k \in \mathbb{R}^{n \times n}$ is a diagonal matrix defined as $\boldsymbol{W}_k = \mathrm{diag}(p_{k;1}, \ldots, p_{k;n})$ with

$$p_{k;i} = \exp([\boldsymbol{X}_k]_{i,\bullet}^{\mathrm{T}} \boldsymbol{\beta}_k)/\{1 + \exp([\boldsymbol{X}_k]_{i,\bullet}^{\mathrm{T}} \boldsymbol{\beta}_k)\}^2 \quad \text{for} \quad i = 1, \ldots, n.$$

Since $\boldsymbol{\ell}_k''(\boldsymbol{\beta}_k)$ does not depend on $\boldsymbol{y}$, we have $\boldsymbol{H}_k = (1/n)\boldsymbol{\ell}_k''(\boldsymbol{\beta}_k)$, for $1 \leqslant k \leqslant |\mathcal{M}|$. By simple algebra, it can be verified that Condition (A1) is satisfied for logistic regression model too.

To estimate the bias of the model averaging estimator, we define $\boldsymbol{\beta}_k^*$ as the solution of the equation $\mathrm{E}[\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k)] = \mathrm{E}\{\boldsymbol{X}_k^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{p}_k)\} = \boldsymbol{0}$, i.e., $\boldsymbol{\beta}_k^*$ is a solution of

$$\boldsymbol{X}_k^{\mathrm{T}} (\boldsymbol{p}_{\mathrm{true}} - \boldsymbol{p}_k) = 0. \tag{3.5}$$

Denote by $\boldsymbol{p}_k^* = \exp(\boldsymbol{X}_k \boldsymbol{\beta}_k^*)/\{1 + \exp(\boldsymbol{X}_k \boldsymbol{\beta}_k^*)\} \in \mathbb{R}^n$ calculated componentwise. We have $\boldsymbol{X}_k^{\mathrm{T}}(\boldsymbol{p}_{\mathrm{true}} - \boldsymbol{p}_k^*) = 0$, and it follows that

$$\begin{aligned}
\mathrm{E}\boldsymbol{\ell}_k'(\boldsymbol{\beta}_k^*)\boldsymbol{\ell}_{k'}'(\boldsymbol{\beta}_k^*)^{\mathrm{T}} &= \boldsymbol{X}_k^{\mathrm{T}} \mathrm{E}(\boldsymbol{y} - \boldsymbol{p}_k^*)(\boldsymbol{y} - \boldsymbol{p}_{k'}^*)^{\mathrm{T}} \boldsymbol{X}_{k'} \\
&= \boldsymbol{X}_k^{\mathrm{T}} \mathrm{E}\{(\boldsymbol{y} - \boldsymbol{p}_{\mathrm{true}}) - (\boldsymbol{p}_k^* - \boldsymbol{p}_{\mathrm{true}})\}\{(\boldsymbol{y} - \boldsymbol{p}_{\mathrm{true}}) - (\boldsymbol{p}_{k'}^* - \boldsymbol{p}_{\mathrm{true}})\}^{\mathrm{T}} \boldsymbol{X}_{k'} \\
&= \boldsymbol{X}_k^{\mathrm{T}} \mathrm{E}(\boldsymbol{y} - \boldsymbol{p}_{\mathrm{true}})(\boldsymbol{y} - \boldsymbol{p}_{\mathrm{true}})^{\mathrm{T}} \boldsymbol{X}_{k'} = \boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{W}^{\mathrm{true}} \boldsymbol{X}_{k'}
\end{aligned}$$

where $\boldsymbol{W}^{\mathrm{true}} = \mathrm{var}(\boldsymbol{y} - \boldsymbol{p}_{\mathrm{true}}) = \mathrm{E}(\boldsymbol{y} - \boldsymbol{p}_{\mathrm{true}})(\boldsymbol{y} - \boldsymbol{p}_{\mathrm{true}})^{\mathrm{T}}$. In addition, write $\boldsymbol{W}_k^* = \mathrm{diag}(\boldsymbol{p}_k^*) \in \mathbb{R}^{n \times n}$.

The gradient ∇p is given by $\nabla p(\boldsymbol{\beta}_k^*) = p_k^*(1-p_k^*)\boldsymbol{x}_k^*$, $1 \leqslant k \leqslant |\mathcal{M}|$. Thus, the MSE estimate is

$$\begin{aligned} Q(\boldsymbol{w}) = &\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'}(p_k^* - p_{\text{true}})(p_{k'}^* - p_{\text{true}}) \\ &+ \sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'} p_k^*(1-p_k^*)\boldsymbol{x}_k^{*\mathrm{T}}\{\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{W}_k^*(\boldsymbol{I}_n - \boldsymbol{W}_k^*)\boldsymbol{X}_k\}^{-1} \\ &\times \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{W}^{\text{true}}\boldsymbol{X}_{k'}\{\boldsymbol{X}_{k'}^{\mathrm{T}}\boldsymbol{W}_{k'}^*(\boldsymbol{I}_n - \boldsymbol{W}_{k'}^*)\boldsymbol{X}_{k'}\}^{-1}\boldsymbol{x}_k^* p_k^*(1-p_k^*). \end{aligned}$$

However, $Q(\boldsymbol{w})$ involves unknown $\boldsymbol{\beta}_{\text{true}}$ and $\boldsymbol{\beta}_k^*$. As in the linear regression model case, we use the full model to estimate $\boldsymbol{\beta}_{\text{true}}$ and denote by the estimator $\widehat{\boldsymbol{\beta}}_{\text{full}}$. Then, compute $\widehat{\boldsymbol{p}}_{\text{full}} = \boldsymbol{p}(\widehat{\boldsymbol{\beta}}_{\text{full}})$ and $\widehat{p}_{\text{full}} = p(\widehat{\boldsymbol{\beta}}_{\text{full}})$. The estimators $\widehat{\boldsymbol{p}}_k^* = \exp(\boldsymbol{X}_k\widehat{\boldsymbol{\beta}}_k)/\{1+\exp(\boldsymbol{X}_k\widehat{\boldsymbol{\beta}}_k)\}$ and $\widehat{p}_k^* = \exp(\boldsymbol{x}_k^{*\mathrm{T}}\widehat{\boldsymbol{\beta}}_k)/\{1+\exp(\boldsymbol{x}_k^{*\mathrm{T}}\widehat{\boldsymbol{\beta}}_k)\}$ are obtained by solving the equation

$$\boldsymbol{X}_k^{\mathrm{T}}(\widehat{\boldsymbol{p}}_{\text{full}} - \boldsymbol{p}_k) = 0, \tag{3.6}$$

using iterative re-weighted least squares (IRLS) method (see, e.g., [14]). Specifically, let $\boldsymbol{\beta}_k{}^{(s)}$ be the solution of (3.6) at the s-th stage of the IRLS algorithm. The coefficients for the $(s+1)$-th stage is then given by

$$\boldsymbol{\beta}_k{}^{(s+1)} = \boldsymbol{\beta}_k^{(s)} + \{\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{W}_k(\boldsymbol{I}_n - \boldsymbol{W}_k)\boldsymbol{X}_k\}^{-1}\boldsymbol{X}_k^{\mathrm{T}}\left\{\frac{\exp(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{full}})}{1+\exp(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{full}})} - \frac{\exp(\boldsymbol{X}_k\boldsymbol{\beta}_k)}{1+\exp(\boldsymbol{X}_k\boldsymbol{\beta}_k)}\right\}\Bigg|_{\boldsymbol{\beta}_k=\boldsymbol{\beta}_k{}^{(s)}},$$

for $s = 0, 1, 2, \ldots$ When the algorithm converges, we obtain the estimate $\widehat{\boldsymbol{\beta}}_k$. Putting together, we estimate $Q(\boldsymbol{w})$ by

$$\begin{aligned} \hat{Q}(\boldsymbol{w}) = &\sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'}(\widehat{p}_k^* - \widehat{p}_{\text{full}})(\widehat{p}_{k'}^* - \widehat{p}_{\text{full}}) \\ &+ \sum_{k\in\mathcal{M}}\sum_{k'\in\mathcal{M}} w_k w_{k'}\widehat{p}_k^*(1-\widehat{p}_k^*)\boldsymbol{x}_k^{*\mathrm{T}}\{\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{W}_k^*(\boldsymbol{I}_n - \boldsymbol{W}_k^*)\boldsymbol{X}_k\}^{-1} \\ &\times \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{W}^{\text{true}}\boldsymbol{X}_{k'}\{\boldsymbol{X}_{k'}^{\mathrm{T}}\boldsymbol{W}_{k'}^*(\boldsymbol{I}_n - \boldsymbol{W}_{k'}^*)\boldsymbol{X}_{k'}\}^{-1}\boldsymbol{x}_k^*\widehat{p}_k^*(1-\widehat{p}_k^*)\,|_{\boldsymbol{p}_k^*=\widehat{\boldsymbol{p}}_k^*;\boldsymbol{p}_{k'}^*=\widehat{\boldsymbol{p}}_{k'}^*;\,\boldsymbol{p}_{\text{true}}=\widehat{\boldsymbol{p}}_{\text{full}}}. \end{aligned} \tag{3.7}$$

We can obtain $w_1, \ldots, w_N$ such that the estimated MSE $\hat{Q}(\boldsymbol{w})$ is minimized, similar to the development done in linear regression setup. These weights can be assigned to individual models for developing the model averaging estimator.

4 Simulation study and real data analysis

4.1 Simulation study I: Bias and variance tradeoff

We study both finite and large sample behavior of the model averaging estimator under a regression setup: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. In the study, $p = 9$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_9)^{\mathrm{T}}$, where β_0 is the intercept coefficient. We assume that 5 parameters $(\beta_0, \ldots, \beta_4)^{\mathrm{T}}$ are always included in all candidate models and the remaining parameters $(\beta_5, \ldots, \beta_9)^{\mathrm{T}}$ may or may not be in a candidate model. For simulation of $\boldsymbol{y}$, first we set the true parameter (henceforth, referred to as $\boldsymbol{\beta}^*$) as follows:

$$\boldsymbol{\beta}^* = (\underbrace{0.3, 0.3, 0.5, 0.1, 0.5}_{\text{always included}}, \underbrace{0.0, 0.6, 0.0, 0.1, 0.0}_{\text{candidate parameters}})^{\mathrm{T}}.$$

For the design matrix, the first column of $\boldsymbol{X}$ is chosen to be 1 (for intercept) and the rest are simulated independently from $\mathsf{N}(0,1)$ distribution. The final response $\boldsymbol{y}$ is obtained by adding independent Gaussian error $\epsilon_i \sim \mathsf{N}(0,1)$ to each row. We also simulate $\boldsymbol{x}^* = (1, x_1^*, \ldots, x_9^*)^{\mathrm{T}}$ so that each element x_j^* is simulated from $\mathsf{N}(0,1)$ and define our parameter of interest $\mu^* = \boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}^*$.

Clearly, based on all possible choices of last 5 parameters, there are a total of $2^5 = 32$ candidate

models. For ease of calculations we will consider the following 6 nested set of candidate models and the true/oracle model (represented pictorially):

$$\begin{array}{c} \\ \textbf{Candiate } 1 \\ \textbf{Candiate } 2 \\ \textbf{Candiate } 3 \\ \textbf{Candiate } 4 \\ \textbf{Candiate } 5 \\ \textbf{Candiate } 6 \\ \textbf{Oracle} \end{array} \begin{array}{c} \begin{array}{ccccc} \boldsymbol{\beta_5} & \boldsymbol{\beta_6} & \boldsymbol{\beta_7} & \boldsymbol{\beta_8} & \boldsymbol{\beta_9} \end{array} \\ \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \times & \checkmark & \checkmark & \checkmark & \checkmark \\ \times & \times & \checkmark & \checkmark & \checkmark \\ \times & \times & \times & \checkmark & \checkmark \\ \times & \times & \times & \times & \checkmark \\ \times & \times & \times & \times & \times \\ \times & \checkmark & \times & \checkmark & \times \end{pmatrix} \end{array}. \tag{4.1}$$

Note that the true model is a sub-model of candidate models 1 and 2, and candidate model 6 only contains the first 5 fixed parameters and none of the candidate parameters are included. We will consider two cases. In Case A we will consider all 7 models in (4.1) comprising of the 6 nested models and the true model. In Case B, we will only consider the first 6 nested models. We will compare our results with that of the *oracle estimate*, where we know before-hand which parameters are non-zero and use a least squares method to estimate $\boldsymbol{\beta}$ and consequently μ^*. We vary the sample size n from 100 to 1,000 and compare the bias and variance between the proposed and the oracle method.

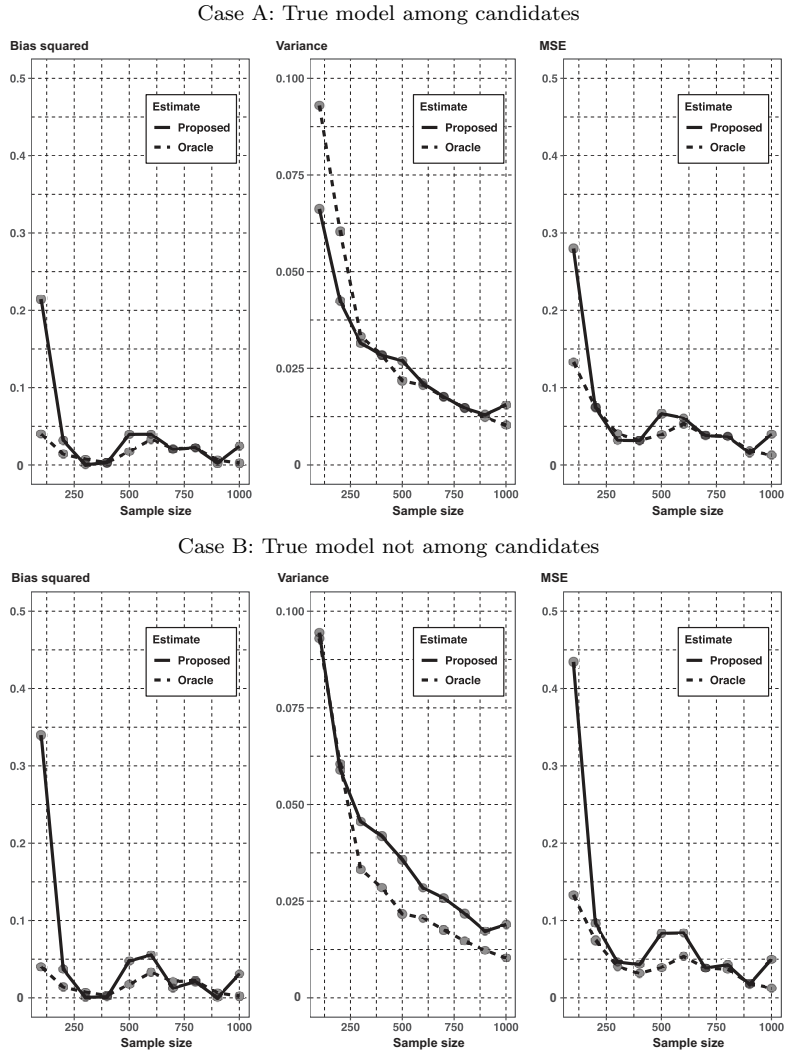

Figure 1 Bias and variance movement for the proposed model averaging and the oracle estimator of μ^*. The true model is a sub-model of (nested within) some of the candidate models, but not included in the candidate model set in Case B

In Figure 1, we consider two cases: Case A where the true (or oracle) model is one of the candidate models and Case B where the true model is not one of the candidate models. In Figure 1 we compare the squared bias, variance and mean squared error movements as the sample size is increased. In the top panel for Case A, the model-average estimator has smaller variance than the oracle estimate even for very small sample sizes which is to be expected; the reason being that the candidate model set contains the oracle as one of its candidates and further averaging reduces variances. In the bottom panel for Case B, with the increase in the sample size, the variance of the proposed estimator decreases but is slightly higher compared to oracle. In both cases, the bias matches the oracle very closely as the sample size increases. It is suggestive from the plots in Figure 1 that in the linear regression setup, even when the candidate models do not include the true set of parameters, model averaging approaches the performance of the oracle estimator in terms of bias and variance. We want to stress that this close performance of the model averaging estimator as compared to oracle is specific to this simple linear regression setup where the true model is a sub-model of some of the candidate models. In general, the question of *whether the performance of the model averaging estimator is close to the oracle*, would require separate investigation specific to the model and data at hand.

4.2 Simulation study II: Comparison with existing model averaging methods

In this subsection we use both linear and logistic regression models to perform simulation studies to compare the performance of the frequentist model averaging estimator with the proposed weights with two existing model averaging methods by Hjort and Claeskens [11] and Liang et al. [20]. The method by Hjort and Claeskens [11] using AIC based weights (which we refer to as the frequentist model averaging (FMA) method) and the method by Liang et al. [20] (which we refer to as the optional weighting (OPT) method) are two well-studied approaches and both are also close to ours. The FMA method combines estimators from different models with the assumption that the data are coming from a local mis-specification framework so the candidate model used has to have a bias of $\mathcal{O}(1/\sqrt{n})$ or less. We do not have this restriction in our proposed method. The OPT method proposes an unbiased estimate of MSE of the model averaging estimator and then the model averaging weights are obtained by minimizing the trace of the MSE estimate. The weight selection for OPT has been shown to exhibit optimality properties in terms of minimizing the MSE. However, their development is limited only to linear regression setting.

Linear regression. In the linear regression setup, we work with a design similar to the one we described in Subsection 4.1. In particular, in the setup $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, we take $p = 4$ and $n = 100$; we denote $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ with β_0 being the coefficient for the intercept. In this setup the fixed parameter is β_0 (i.e., $k = 1$) and the rest may or may not appear in the model (i.e., $m = 3$). As before, we use $\boldsymbol{\beta}^*$ to denote the true parameter. The elements of the design matrix $\boldsymbol{X}$ are simulated independently from an $\mathsf{N}(0, 1)$ distribution and the elements of the error vector $\boldsymbol{\varepsilon}$ are simulated independently as $\mathsf{N}(0, 1)$.

In this simulation setup, the estimand of interest is the following:

$$\mu^* = \boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}^*, \quad \text{where} \quad \boldsymbol{x}^* \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{I}_4).$$

For our specific example, we have $\boldsymbol{x}^* = (1, -1.855445, -1.018565, -1.045111)$ and the true parameter $\boldsymbol{\beta}^* = (0.3, 0.1, 0.3, \beta_3^*)$. In the following we will consider two cases as before: Case A, where we will vary the value of β_3^* in the set $\{0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ and Case B, where we will vary β_3^* in the set $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. Note that we have not included the case $\beta_3^* = 0$ for Case B in order to make sure that the true model is not included among the candidates. As before, we will consider two different sets of candidate models:

$$\begin{aligned} &\textbf{Case A.} \quad \{\beta_0\},\ \{\beta_0, \beta_1\},\ \{\beta_0, \beta_1, \beta_2\},\ \{\beta_0, \beta_1, \beta_2, \beta_3\};\\ &\textbf{Case B.} \quad \{\beta_0\},\ \{\beta_0, \beta_1\},\ \{\beta_0, \beta_1, \beta_2\}. \end{aligned} \tag{4.2}$$

Table 1 (Linear regression) Mean squared error for estimation of μ^* for the (a) model averaging estimator with proposed weights, (b) model averaging estimator with Liang's weights [20], (c) Hjort's model [11] averaging estimator with AIC based weights, and (d) oracle estimator. Here, in the top table, the candidate models include the true set of parameters (Case A) and in the bottom table the true set of parameters is not included (Case B)—as described in (4.2)

Case A. True model among candidates									
		(a) Proposed		(b) OPT		(c) FMA		(d) Oracle	
β_3^*	μ^*	Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
0	−0.191	−0.164	0.333	−0.070	0.313	0.167	0.384	−0.236	0.315
0.001	−0.192	−0.164	0.333	−0.071	0.313	0.167	0.384	−0.241	0.334
0.005	−0.196	−0.167	0.333	−0.075	0.314	0.165	0.387	−0.245	0.334
0.010	−0.202	−0.172	0.335	−0.079	0.315	0.164	0.391	−0.251	0.334
0.050	−0.243	−0.211	0.339	−0.132	0.319	0.154	0.420	−0.292	0.334
0.100	−0.296	−0.259	0.344	−0.162	0.339	0.146	0.461	−0.345	0.334
0.500	−0.714	−0.697	0.343	0.064	0.803	0.128	0.851	−0.763	0.334
Case B. True model not among candidates									
		(a) Proposed		(b) OPT		(c) FMA		(d) Oracle	
β_3^*	μ^*	Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
0.001	−0.192	−0.143	0.330	0.038	0.336	0.206	0.416	−0.241	0.334
0.005	−0.196	−0.145	0.331	0.038	0.339	0.205	0.420	−0.245	0.334
0.010	−0.202	−0.148	0.331	0.039	0.342	0.205	0.424	−0.251	0.334
0.050	−0.243	−0.168	0.334	0.041	0.371	0.199	0.458	−0.292	0.334
0.100	−0.296	−0.180	0.341	0.043	0.412	0.191	0.501	−0.345	0.334
0.500	−0.714	−0.134	0.654	0.063	0.802	0.128	0.852	−0.763	0.334

Table 2 (Logistic regression) Estimation of p^* for the (a) model averaging estimator with proposed weights, (b) FMA: Hjort's model [11] averaging estimator with AIC based weights, and (c) oracle estimator. Here, in the top table, the candidate models include the true set of parameters (Case A) and in the bottom table the true set of parameters is not included (Case B)—as described in (4.2)

Case A. True model among candidates							
		(a) Proposed		(b) FMA		(c) Oracle	
β_3^*	p^*	Estimate	Error	Estimate	Error	Estimate	Error
0	0.452	0.469	0.130	0.510	0.106	0.436	0.127
0.001	0.452	0.468	0.131	0.510	0.105	0.443	0.144
0.005	0.451	0.468	0.131	0.510	0.106	0.443	0.144
0.010	0.450	0.466	0.133	0.510	0.107	0.441	0.145
0.050	0.439	0.454	0.129	0.506	0.110	0.424	0.139
0.100	0.427	0.444	0.132	0.505	0.119	0.412	0.141
0.500	0.329	0.371	0.144	0.521	0.208	0.331	0.133
Case B. True model not among candidates							
		(a) Proposed		(b) FMA		(c) Oracle	
β_3^*	p^*	Estimate	Error	Estimate	Error	Estimate	Error
0.001	0.452	0.475	0.124	0.528	0.109	0.443	0.144
0.005	0.451	0.475	0.124	0.529	0.110	0.443	0.144
0.010	0.450	0.473	0.126	0.529	0.111	0.441	0.145
0.050	0.439	0.463	0.121	0.530	0.121	0.424	0.139
0.100	0.427	0.464	0.127	0.531	0.131	0.412	0.141
0.500	0.329	0.466	0.170	0.533	0.218	0.331	0.133

Note that in Case A, the true parameter set is included in the model while in Case B, the true parameter set is not included. In fact, Case B represents a typical scenario where the researcher is not even aware of the presence of the existence of the feature corresponding to β_3 and hence is working under a mis-specified model.

In Table 1 the performances of different methods are compared for Case A (at the top) and Case B (bottom). For each separate choice of β_3^*, we performed 100 simulations and reported their averages in Table 1 along with the root mean squared error. Specifically the error for this simulation setup was

defined as,

$$\text{Error} = \sqrt{(1/100)\sum_{k=1}^{100} |\widehat{\mu}_k - \mu^*|^2},$$

where $\widehat{\mu}_k$ is the estimate corresponding to a specific method at the k-th simulation. In Case A of Table 1, we compare the methods when β_3 is included in the largest candidate model while in Case B, β_3 is not considered in any of the candidate models. From Case A of Table 1, it can be seen that in the finite sample framework ($n = 100$), the performances of the proposed model-average estimator and OPT are similar and both outperform FMA. Moreover with the increase in magnitude of β_3^* to 0.5, the proposed model averaging method outperforms both FMA and OPT. On the other hand, the setup in Case B of Table 1 shows that with the increase in β_3^*, the estimation error increases consistently for all three methods. Nevertheless, our proposed method clearly outperforms the competing methods in this scenario for all β_3^* values. We also remark that the proposed method performs well up till $\beta_3 = 0.1$, but the error jumps for the larger signal with $\beta_3 = 0.5$. This is expected since β_3 is not considered in any of the candidate models and the extent of model mis-specification is large at $\beta_3^* = 0.5$.

Logistic regression. We now describe the efficacy of the proposed methodology for the logistic regression setup and compare its performance with Hjort's FMA method with AIC-based weights [11]. The logit model is given by

$$p_i = \mathrm{P}(y_i = 1 \,|\, \boldsymbol{X}) = \frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}, \quad \forall\, i = 1, \ldots, n, \tag{4.3}$$

where

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}, \quad \boldsymbol{x}_i \in \mathbb{R}^p \quad \text{and} \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

We take $n = 100$ and $p = 4$ where the intercept is always included ($k = 1$) and the rest of the parameters can be varied in forming candidate models ($m = 3$). As in the linear regression simulation setup, the elements of $\boldsymbol{X}$ are simulated independently from $\mathsf{N}(0, 1)$ distribution. In this setup, the true value of the parameter $\boldsymbol{\beta}$ is set as $\boldsymbol{\beta}^* = (0.3, 0.1, 0.3, \beta_3^*)$. As in the linear regression setup, we consider two cases namely, Case A and Case B; see (4.2) for more details. We vary the value of β_3^* (as before) in the set $\{0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ for Case A and $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ for Case B. For this logistic regression setup, our estimand of interest is as follows:

$$p^* = \exp(\eta^*)/(1 + \exp(\eta^*)) \quad \text{where} \quad \eta^* = \boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\beta}^* \quad \text{and} \quad \boldsymbol{x}^* \sim \mathsf{N}_4(\boldsymbol{0}, \boldsymbol{I}_4). \tag{4.4}$$

As in the regression setup, we set

$$\boldsymbol{x}^* = [1.0, -1.86, -1.019, -1.045].$$

Note that the specifics of model averaging estimator for the estimand in (4.4) has been described in detail in Subsection 3.2. Specifically, (3.7) describes the MSE function to be minimized for optimal weights. We compare our prposed method with Hjort's FMA method with AIC-based weights [11] and the oracle estimate. The results for both Cases A and B are summarized in Table 2. We define the error metric, based on 100 simulations, as,

$$\text{Error} = \sqrt{(1/100)\sum_{k=1}^{100} |\widehat{p}_k - p^*|^2},$$

where $\widehat{p}_k$ is the estimate corresponding to a specific method at the k-th simulation. As in the linear regression setup, for the logistic regression as well, we see that the proposed method performs better than Hjort's method using AIC-based weights in both cases across all β_3^* values. For Case A, the performance of our proposed method matches that of the oracle and the differences are within the margin of error. For Case B, the performance of our proposed method tracks the oracle well until the signal strength of β_3^* is increased to 0.5, in which case the estimation error increases.

Table 3 Prediction error for different methods for prostate cancer data

Method used	test error
Model selection (best subset regression)	0.487
Model averaging (proposed weights)	0.453
Model averaging (AIC weights)	0.987
Full model	1.272

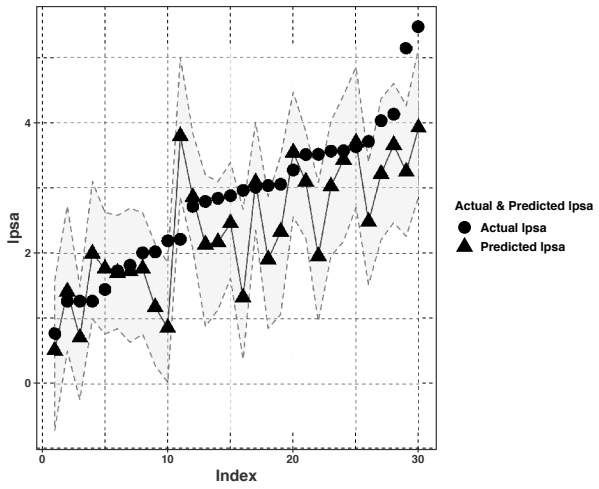

Figure 2 Actual and predicted level of lpsa (level of prostate-specific antigen) based on the prostate cancer data from [28]. In the x-axis the indices of the 30 observations are noted. In the y-axis we note the lpsa values. The circle points indicate actual (observed) values while the triangle points indicate predicted values based on average of 50 replications. The gray band denotes the 90% confidence interval

4.3 Analysis of prostate cancer data

The data for this example come from a study by Stamey et al. [28]. They examined the relationship between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. As a regression problem, the response variable is lpsa, the level of prostate-specific antigen, with values ranging from -0.43 to 5.58. The predictor variables (clinical measures) are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45). Here, svi is a binary variable, and gleason is an ordered categorical variable.

As the response variable lpsa is continuous, we have used a standard linear regression model as the base method for applying proposed model averaging procedure which was developed in Subsection 3.1. We considered 8 nested and an intercept-only model as candidate models for applying the proposed (as well as competing) model averaging methods. The candidate models are shown pictorially in (4.5) below:

$$\begin{array}{l} \\ \textbf{Candiate } 1 \\ \textbf{Candiate } 2 \\ \textbf{Candiate } 3 \\ \textbf{Candiate } 4 \\ \textbf{Candiate } 5 \\ \textbf{Candiate } 6 \\ \textbf{Candiate } 7 \\ \textbf{Candiate } 8 \\ \textbf{Candiate } 9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} \text{Intercept} & \text{age} & \text{gleason} & \text{lcp} & \text{svi} & \text{lbph} & \text{pgg45} & \text{lweight} & \text{lcavol} \end{array} \\ \begin{pmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \times & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \times & \times & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \times & \times & \times & \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \times & \times & \times & \times & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \times & \times & \times & \times & \times & \checkmark & \checkmark & \checkmark \\ \checkmark & \times & \times & \times & \times & \times & \times & \checkmark & \checkmark \\ \checkmark & \times & \times & \times & \times & \times & \times & \times & \checkmark \\ \checkmark & \times & \times & \times & \times & \times & \times & \times & \times \end{pmatrix} \end{array}. \tag{4.5}$$

We considered a best-subset model selection approach using an all-subsets search. In this model selection approach, the estimated prediction error is obtained using a crude cross-validation method: the

dataset is divided randomly into a training set of size 67 and a test set of size 30. The training set is used to select a model and then the test set is used to compute the prediction error, averaging over all 30 points. We repeat the process five times and average over the five prediction errors. We also considered the model averaging method using two different sets of weights: the proposed weights and also AIC-based weights. Using the proposed weights, the proposed approach assigned the most weights to the model with features lcavol, lweight, svi, pgg45, lcp, gleason and lbph and the model with lcavol, lweight, svi, pgg45, lcp, gleason, lbph and age. The procedure with AIC-based weights gives more weight to a smaller model containing lcavol and lweight. We used the same crude cross-validation method as in the best-subset model selection approach, with a training set of size 67 and a test set of size 30. The training set is used to obtain the model averaging estimates and then the test set is used to compute the prediction error, averaging over all 30 points. We repeat the process five times and average over the five prediction errors. Table 3 summarizes the numerical results using the above methods, plus a regression analysis including all covariate variables (full model). It shows that the model averaging method using the proposed weights has the smallest empirical average prediction error on the testing samples among all methods.

Finally, as an illustration, we also plotted in Figure 2, a set of 90% prediction intervals of antigen levels for one test dataset in one of our simulation runs. The x-axis is the index of the 30 observations in the test dataset. In order to get the prediction interval, we kept the test dataset fixed while in 50 different replications we selected a random subset of 50 observations from the training data (of original size 67) and applied the model averaging method to analyze the training data of size 50 and use the result to predict the lpsa values for the test dataset. In order to construct the prediction interval we added to each predicted mean, a Gaussian noise with mean 0 and standard error equal to the estimated standard error from the full model denoted as $\widehat{\sigma}_{\text{full}}$ (see (3.3)). In this case the estimate was $\widehat{\sigma}_{\text{full}} = 0.599$. The upper and lower limits of the prediction band were calculated based on quantiles. As is clear from the plots, most of the observations fall within the 90% prediction interval.

5 Discussion

In this paper, we propose a more general framework where the choice of true model is not fixed. The truth can be any one or a mixture of the candidate models. Models that have large biases are not excluded from our analysis. We study the behavior of frequentist model averaging estimator with an optimal weighting scheme to combine all the individual candidate models. As an illustration, we derive the model averaging estimator in the linear and logistic regression framework. We also implement the weighting scheme proposed by Liang et al. [20] and compare their performance to the FMA method with AIC based weights as in [11]. The simulation results indicate that under certain model specifications, the proposed estimator works better than FMA estimator with AIC-weights from [11].

There are many ways a regression model can be mis-specified. Mis-specification in most cases is often interpreted as a case of left out variables or when the functional form of the model is not correctly specified. In these instances, the normality assumption among random errors are violated. This results in the estimates being biased as discussed in [8]. These estimates can harm the decision making process, so one should be very attentive while fitting and choosing models in the presence of mis-specification. Many methods have been used to measure and limit mis-specification in model fitting. Ramsey regression equation specification error test, discussed in [29], may help provide a test that is useful in a linear regression setup.

In model averaging, if the true model is not included in the set of candidate models, we end up using an estimate that is biased. If all the models are mis-specified, the weights derived by AIC or by using a consistent or unbiased estimator of mean squared error are not optimal and should be with care. When the true model is not included in the analysis thus all the candidate models are wrong, there have been developments in model selection that takes care of the bias resulting from selection (see [15,16]). Penalized versions of AIC and BIC have been derived that perform better than other selection criteria. One can follow a similar path and derive the model averaging weights based on a slightly modified criterion. Zhang et al. [35] derived a KL-based criterion and proposed a model averaging estimator, which is proved to

be asymptotically optimal (see also [7,34,36] for the related works). However, it is unclear whether their strategies can be used in our framework and this warrants a further study.

Another problem with model averaging is that the number of optional parameters in analysis could be very high. For example, if there are 30 parameters we could end up using as many as 2^{30} candidate models. This may be time consuming and not ideal in certain fields of study. However, as suggested in this paper, a statistician can choose to use all or very few candidate models as per the scope of the study. This could be explored in further developments.

Acknowledgements The work was supported by National Science Foundation of USA (Grant Nos. DMS-1812048, DMS-1737857, DMS-1513483 and DMS-1418042) and National Natural Science Foundation of China (Grant No. 11529101). This article is a work developed based on the thesis of the first author. The authors wish to use this article to celebrate Professor Lincheng Zhao's 75th birthday and his tremendous and long lasting contribution to statistical research and education in China and around the world. The authors also thank two referees for their valuable suggestions and comments that have helped improve the paper substantially.

References

1 Billingsley P. Probability and Measure. Chichester: John Wiley & Sons, 2008
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, Hjort N L. Model selection and model averaging. J Math Psych, 2008, 44: 92–107
4 Danilov D, Magnus J R. Forecast accuracy after pretesting with an application to the stock market. J Forecast, 2004, 23: 251–274
5 Danilov D, Magnus J R. On the harm that ignoring pretesting can cause. J Econometrics, 2004, 122: 27–46
6 Draper D. Assessment and propagation of model uncertainty. J R Stat Soc Ser B Stat Methodol, 1995, 57: 45–97
7 Gao Y, Zhang X, Wang S, et al. Model averaging based on leave-subject-out cross-validation. J Econometrics, 2016, 192: 139–151
8 Giles D E A, Lieberman O, Giles J A. The optimal size of a preliminary test of linear restrictions in a misspecified regression model. J Amer Statist Assoc, 1992, 87: 1153–1157
9 Hansen B E. Least squares model averaging. Econometrica, 2007, 75: 1175–1189
10 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag, 2009
11 Hjort N L, Claeskens G. Frequentist model average estimators. J Amer Statist Assoc, 2003, 98: 879–899
12 Hjort N L, Claeskens G. Focused information criteria and model averaging for the Cox hazard regression model. J Amer Statist Assoc, 2006, 101: 1449–1464
13 Hoeting J A, Madigan D, Raftery A E, et al. Bayesian model averaging. Statist Sci, 1999, 14: 121–149
14 Holland P W, Welsch R E. Robust regression using iteratively reweighted least-squares. Comm Statist Theory Methods, 2007, 6: 813–827
15 Hurvich C M, Tsai C-L. Regression and time series model selection in small samples. Biometrika, 1989, 76: 297–307
16 Hurvich C M, Tsai C-L. Bias of the corrected AIC criterion for underfitted regression and time series models. Biometrika, 1991, 78: 499–509
17 Karagrigoriou A, Lee S, Mattheou K. A model selection criterion based on the BHHJ measure of divergence. J Statist Plann Inference, 2009, 139: 228–235
18 Lehmann E L. Elements of Large-Sample Theory. Springer Texts in Statistics. New York: Springer-Verlag, 1999
19 Lehmann E L, Casella G. Theory of Point Estimation, 2nd ed. Springer Texts in Statistics. New York: Springer-Verlag, 1998
20 Liang H, Zou G, Wan A T K, et al. Optimal weight choice for frequentist model average estimators. J Amer Statist Assoc, 2011, 106: 1053–1066
21 Lien D, Shrestha K. Estimating the optimal hedge ratio with focus information criterion. J Futures Markets, 2005, 25: 1011–1024
22 Madigan D, Raftery A E, York J C, et al. Strategies for graphical model selection. In: Selecting Models from Data: Artificial Intelligence and Statistics IV. Lecture Notes in Statistics, vol. 89. New York: Springer, 1994, 91–100
23 Magnus J R, Wan A T K, Zhang X. Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. Comput Statist Data Anal, 2011, 55: 1331–1341
24 Mitra P. Topics in model averaging & toxicity models in combination therapy. PhD Thesis. New Brunswick: Rutgers University, 2015

25 Pesaran M H, Schleicher C, Zaffaroni P. Model averaging in risk management with an application to futures markets. J Empir Finance, 2009, 16: 280–305
26 Posada D, Buckley T R. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology, 2004, 53: 793–808
27 Raftery A E, Madigan D, Hoeting J A. Bayesian model averaging for linear regression models. J Amer Statist Assoc, 1997, 92: 179–191
28 Stamey T A, Kabalin J N, Ferrari M, et al. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, IV: Anti-androgen treated patients. J Urol, 1989, 141: 1088–1090
29 Thursby J G, Schmidt P. Some properties of tests for specification error in a linear regression model. J Amer Statist Assoc, 1977, 72: 635–641
30 Van der Vaart A W. Asymptotic Statistics, Volume 3. Cambridge: Cambridge University Press, 2000
31 Wan A T K, Zhang X, Zou G. Least squares model averaging by Mallows criterion. J Econometrics, 2010, 156: 277–283
32 Wei Y, McNicholas P D. Mixture model averaging for clustering and classification. Adv Data Anal Classif, 2015, 22: 197–217
33 Zhang X, Wan A T K, Zhou S Z. Focused information criteria, model selection and model averaging in a tobit model with a non-zero threshold. J Bus Econom Statist, 2012, 30: 132–142
34 Zhang X, Yu D, Zou G, et al. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. J Amer Statist Assoc, 2016, 111: 1775–1790
35 Zhang X, Zou G, Carroll R J. Model averaging based on Kullback-Leibler distance. Statist Sinica, 2015, 25: 1583–1598
36 Zhang X, Zou G, Liang H. Model averaging and weight choice in linear mixed-effects models. Biometrika, 2014, 101: 205–218

Appendix A

Appendix A.1 Regularity conditions and assumptions

In this appendix we state the regularity conditions that were used throughout the paper. We assume that the density function satisfies the following conditions:

(a) Θ is an open subset of $\mathbb{R}^p$, and the support of the density $f(y,\boldsymbol{\beta})$ is independent of $\boldsymbol{\beta}$.

(b) The true parameter value is an interior point of the parameter space.

(c) $\ell'_{k;i}$ and $\ell''_{k;i}(\boldsymbol{\beta}^*_k)$ exist and $\ell'_{k;i}$ is a continuous function of $\boldsymbol{\beta}$.

(d) $\mathrm{E}[\ell'_{k;i}]=0$ and $\mathrm{E}[\ell'_{k;i}\ell'^{\mathrm{T}}_{k;i}]=-\mathrm{E}[\ell''_{k;i}(\boldsymbol{\beta}^*_k)]$. These conditions are standard conditions for asymptotic normality of maximum likelihood estimators.

(e)

$$\lim_{n\to\infty}\frac{1}{n}[\ell''_k(\boldsymbol{\beta}^*_k)]=\boldsymbol{H}_k$$

and $\boldsymbol{H}_k$ is positive definite.

(f) For some $\epsilon>0$,

$$\sum_i \mathrm{E}|\lambda'\ell'_{k;i}(\boldsymbol{\beta}_{\mathrm{true}})|^{2+\epsilon}/n^{(2+\epsilon)/2}\to 0 \quad \text{for all} \quad \epsilon\in\mathbb{R}.$$

(g) There exist $\epsilon>0$ and random variables $B_i(y_i)$,

$$\sup\{|\ell''_{k;i}(\boldsymbol{\beta}^*_k)| : \|t-\boldsymbol{\beta}_{\mathrm{true}}\|\leqslant\epsilon\}\leqslant B_i(y_i)$$

and $\mathrm{E}|B_i(y_i)|^{1+\delta}\leqslant K$, where δ and K are positive constants.

We also assume that the variance matrix of the score statistic is finite and positive definite.

Consider a functional $\mu:\mathbb{R}^{p+q}\to\mathbb{R}$. Define $\mu^{(\mathrm{drop})}:\mathbb{R}^{p+m}\to\mathbb{R}$ as the same function as μ with only the $(q-m)$ corresponding arguments dropped. For any $\boldsymbol{b}=(b_1,\ldots,b_p,b_{p+1},\ldots,b_{p+m})$ with $1\leqslant m\leqslant q$ define the $\boldsymbol{c}$-*augmented* version of $\boldsymbol{b}$ as $\widetilde{\boldsymbol{b}}=\{\boldsymbol{b}^{\mathrm{T}},\boldsymbol{c}^{\mathrm{T}}\}^{\mathrm{T}}\in\mathbb{R}^{p+q}$ with some fixed $\boldsymbol{c}\in\mathbb{R}^{q-m}$ inserted at the place of missing components. Let the indices of the missing components be $\{p+i_1,\ldots,p+i_{q-m}\}$. We define $\widetilde{\mu}:\mathbb{R}^{p+m}\to\mathbb{R}$ as the restriction of $\mu:\mathbb{R}^{p+q}\to\mathbb{R}$ subject to $b_{p+i_1}=c_1,\ldots,b_{p+i_{q-m}}=c_{q-m}$. Clearly then $\mu(\widetilde{\boldsymbol{b}})=\widetilde{\mu}(\boldsymbol{b})$. Given a function μ, the fixed value $\boldsymbol{c}$ is chosen in such a way that $\mu(\widetilde{\boldsymbol{b}})=\mu^{(\mathrm{drop})}(\boldsymbol{b})$. We assume that $\boldsymbol{\mu}:\mathbb{R}^{p+q}\to\mathbb{R}^{\ell}$ is a function that is 1st order partially differentiable at $\boldsymbol{\beta}_{\mathrm{true}}$. Note that

by definition of $\boldsymbol{c}$-augmentation, $\mu(\widetilde{\boldsymbol{\beta}}_k) = \mu^{(\mathrm{drop})}(\widehat{\boldsymbol{\beta}}_k)$. For ease of reading, in the subsequent proof, we omit the superscript '(drop)'.

Appendix A.2 Proof of Theorem 2.1

From usual regularity conditions on the log-likelihood, it can be shown that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*) = -\boldsymbol{H}_k^{-1}\bigg\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'_{k;i}(\boldsymbol{\beta}_k^*)\bigg\} + o_{\mathrm{P}}(1).$$

For more details and exact conditions, see [30, Chapter 5].

Now by application of Taylor expansion,

$$\mu(\widehat{\boldsymbol{\beta}}_k) - \mu(\boldsymbol{\beta}_k^*) = \nabla\mu(\boldsymbol{\beta}_k^*)^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*) + o_{\mathrm{P}}(\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|),$$

so that

$$\sqrt{n}(\mu(\widehat{\boldsymbol{\beta}}_k) - \mu(\boldsymbol{\beta}_k^*)) = -\nabla\mu(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\bigg[\boldsymbol{H}_k^{-1}\bigg\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'_{k;i}(\boldsymbol{\beta}_k^*)\bigg\} + o_{\mathrm{P}}(1)\bigg] + o_{\mathrm{P}}(\sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|).$$

Thus it follows that for $0 \leqslant w_k \leqslant 1$ with $\sum_{k\in\mathcal{M}} w_k = 1$,

$$\begin{aligned}
&\sqrt{n}\sum_{k\in\mathcal{M}} w_k\{\mu(\widehat{\boldsymbol{\beta}}_k) - \mu(\boldsymbol{\beta}_{\mathrm{true}})\}\\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k\{\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{\mathrm{true}})\} + \sqrt{n}\sum_{k\in\mathcal{M}} w_k\{\mu(\widehat{\boldsymbol{\beta}}_k) - \mu(\boldsymbol{\beta}_k^*)\}\\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k\{\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{\mathrm{true}})\} - \sum_{k\in\mathcal{M}} w_k\nabla\mu(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\boldsymbol{H}_k^{-1}\bigg\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'_{k;i}(\boldsymbol{\beta}_k^*)\bigg\}\\
&\quad + o_{\mathrm{P}}\bigg(\sum_{k\in\mathcal{M}}\sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|\bigg)\\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k\{\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{\mathrm{true}})\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigg\{-\sum_{k\in\mathcal{M}} w_k\nabla\mu(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\boldsymbol{H}_k^{-1}\ell'_{k;i}(\boldsymbol{\beta}_k^*)\bigg\}\\
&\quad + o_{\mathrm{P}}\bigg(\sum_{k\in\mathcal{M}}\sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|\bigg)\\
&= \sqrt{n}\sum_{k\in\mathcal{M}} w_k\{\mu(\boldsymbol{\beta}_k^*) - \mu(\boldsymbol{\beta}_{\mathrm{true}})\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i + o_{\mathrm{P}}\bigg(\sum_{k\in\mathcal{M}}\sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\|\bigg),
\end{aligned}$$

where we have used the definition that

$$Z_i = -\sum_{k\in\mathcal{M}} w_k\nabla\mu(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\boldsymbol{H}_k^{-1}\ell'_{k;i}(\boldsymbol{\beta}_k^*).$$

First, note that $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*\| = O_{\mathrm{P}}(1)$ via root-n consistency of MLE under our assumptions [30]. Note that Z_i's are independent and $\mathrm{E}Z_i = 0$. Now fix $\epsilon > 0$. In order to prove the asymptotic normality of the quantity $(1/\sqrt{n})\sum_i Z_i$ we invoke the Lindeberg-Feller central limit theorem [1]. This requires verification of the so-called Lindeberg's condition, given by

$$(1/n)\sum_{i=1}^{n}\mathrm{E}Z_i^2\mathbb{I}\{|Z_i| > \sqrt{n}\epsilon\}.$$

Let us denote $Y_{ki} = \nabla\mu(\boldsymbol{\beta}_k^*)\boldsymbol{H}_k^{-1}\ell'_{k;i}$. Now,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}Z_i^2\mathbb{I}\{|Z_i|>\sqrt{n}\epsilon\} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\underbrace{\Big(\sum_{k\in\mathcal{M}}w_kY_{ki}\Big)^2}_{=A,\text{ say}}\underbrace{\mathbb{I}\Big\{\Big|\sum_{k\in\mathcal{M}}w_kY_{ki}\Big|>\sqrt{n}\epsilon\Big\}}_{=B,\text{ say}}\\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big[\sum_{k\in\mathcal{M}}w_kY_{ki}^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}|Y_{ki}|>\sqrt{n}\epsilon\Big\}\Big]\\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big[\max_{k\in\mathcal{M}}|Y_{ki}|^2\mathbb{I}\Big\{\max_{k\in\mathcal{M}}|Y_{ki}|>\sqrt{n}\epsilon\Big\}\Big].\end{aligned}$$

Here, the inequality in the second line is derived by first noting that if $A, B > 0$ and $A < C, B < D$, then $AB < CD$. Secondly, note that $A = (\sum_{k\in\mathcal{M}} w_kY_{ki}) \leqslant \sum_{k\in\mathcal{M}} w_kY_{ki}^2$ by Jensen's inequality. Also since

$$\sqrt{n}\epsilon < \Big|\sum_{k\in\mathcal{M}}w_kY_{ki}\Big| \leqslant \max_{k\in\mathcal{M}}\sum_k|w_k| = 1,$$

it follows that

$$\mathbb{I}\Big\{\Big|\sum_{k\in\mathcal{M}}w_kY_{ki}\Big|>\sqrt{n}\epsilon\Big\} \leqslant \mathbb{I}\Big\{\max_{k\in\mathcal{M}}|Y_{ki}|>\sqrt{n}\epsilon\Big\}.$$

Now take $C = \sum_{k\in\mathcal{M}} w_kY_{ki}^2$ and $D = \mathbb{I}\{\max_{k\in\mathcal{M}}|Y_{ki}| > \sqrt{n}\epsilon\}$.

Now by Condition (A1), the Lindeberg-Feller condition is satisfied for $(1/\sqrt{n})Z_i$'s whence it follows that $(1/\sqrt{n})\sum_{i=1}^n Z_i \sim \mathsf{N}(0,\sigma_w^2)$, where σ_w^2 is given by

$$\sigma_w^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big\{\sum_k w_k\nabla\mu(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\boldsymbol{H}_k^{-1}\ell'_{k;i}\Big\}^2.$$

The theorem follows.

Appendix A.3 Proof of Corollary 2.3

Here, we show that the asymptotic variances are exactly the same for the two results. When we say "asymptotic variance" we mean the first order term in the derived quantities, i.e., when $\sqrt{n}$ is multiplied to the estimator, we ignore all the $o(1)$ terms.

As defined before, for the k-th candidate model, let $\boldsymbol{\beta}_k^* \in \mathbb{R}^{p+|M_k|}$ be the solution of the equation $\mathrm{E}S_k(\boldsymbol{\beta}) = 0$, where $S_k(\boldsymbol{\beta})$ is the score function for the k-th model. Let $\boldsymbol{\beta}_{0,k} = (\boldsymbol{\theta}_0^{\mathrm{T}}, \pi_k\boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+|M_k|}$. Therefore, $\mathrm{E}\{\ell'_k(\boldsymbol{\beta}_k^*)\} = \mathbf{0}$. Then, by Taylor's theorem and appropriate regularity conditions on the density function, it follows that asymptotically, $\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_{0,k} \approx \boldsymbol{J}_k^{-1}\mathrm{E}\{\ell'_k(\boldsymbol{\beta}_0)\}$. Now note that following [11, p. 37],

$$\mathrm{E}\{\ell'_k(\boldsymbol{\beta}_0)\} = \begin{pmatrix}\boldsymbol{J}_{01}\boldsymbol{\delta}/\sqrt{n} + o(1/\sqrt{n})\\ \pi_k\boldsymbol{J}_{11}\boldsymbol{\delta}/\sqrt{n} + o(1/\sqrt{n})\end{pmatrix},$$

so that,

$$\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_{0,k} \approx \boldsymbol{J}_k^{-1}\begin{pmatrix}\boldsymbol{J}_{01}\boldsymbol{\delta}/\sqrt{n}\\ \pi_k\boldsymbol{J}_{11}\boldsymbol{\delta}/\sqrt{n}\end{pmatrix}. \tag{A.1}$$

In order to prove the corollary, we first match the bias terms. Note that in Theorem 2.1, the bias term is given by

$$\sqrt{n}\sum_{k\in\mathcal{M}}w_k\{\mu(\boldsymbol{\beta}_k^*,\gamma_{0,k^c}) - \mu(\boldsymbol{\beta}_{\mathrm{true}})\}.$$

Thus consider term by term, the bias of the k-th component is given by

$$\begin{aligned}\sqrt{n}\{\mu(\boldsymbol{\beta}_k^*,\boldsymbol{\gamma}_{0,k^c})-\mu(\boldsymbol{\beta}_{\rm true})\}&=\sqrt{n}\{\mu(\boldsymbol{\beta}_k^*,\boldsymbol{\gamma}_{0,k^c})-\mu(\boldsymbol{\beta}_0)\}-\sqrt{n}\{\mu(\boldsymbol{\beta}_{\rm true})-\mu(\boldsymbol{\beta}_0)\}\\&\approx\sqrt{n}(\boldsymbol{\beta}_k^*-\boldsymbol{\beta}_{0,k})^{\rm T}\begin{pmatrix}\partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\theta}\\ \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\gamma}_k\end{pmatrix}-\left(\frac{\partial\mu(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\gamma}}\right)^{\rm T}\boldsymbol{\delta}\\&=\begin{pmatrix}\partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\theta}\\ \partial\mu(\boldsymbol{\beta}_0)/\partial\boldsymbol{\gamma}_k\end{pmatrix}^{\rm T}\boldsymbol{J}_k^{-1}\begin{pmatrix}\boldsymbol{J}_{01}\boldsymbol{\delta}\\ \pi_k\boldsymbol{J}_{11}\boldsymbol{\delta}\end{pmatrix}-\left(\frac{\partial\mu(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\gamma}}\right)^{\rm T}\boldsymbol{\delta},\end{aligned}$$

where the last term follows from (A.1). This matches the bias term in (2.7). Looking at the rest, note that from (2.3), the k-th term is given by $\{\nabla\mu(\boldsymbol{\beta}_k^*,\boldsymbol{\gamma}_{0,k^c})\}^{\rm T}\boldsymbol{H}_k^{-1}(\sum_{i=1}^n\ell'_{k;i}(\boldsymbol{\beta}_k^*)/\sqrt{n})+o_P(1)$ (see also the proof of Theorem 2.1). From (A.1), via Taylor's theorem it follows that $\nabla\mu(\boldsymbol{\beta}_k^*,\boldsymbol{\gamma}_{0,k^c})\approx\nabla\mu(\boldsymbol{\beta}_0)$. Also note that from standard theory of maximum likelihood estimation,

$$\begin{aligned}\boldsymbol{H}_k^{-1}(\boldsymbol{\beta}_k^*)\left(\sum_{i=1}^n\ell'_{k;i}(\boldsymbol{\beta}_k^*)\Big/\sqrt{n}\right)&\approx\sqrt{n}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^*)\\&=\sqrt{n}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_{0,k})-\sqrt{n}(\boldsymbol{\beta}_k^*-\boldsymbol{\beta}_{0,k})\\&=\boldsymbol{J}_k^{-1}\begin{pmatrix}\sqrt{n}\overline{U}_n\\ \sqrt{n}\overline{V}_{n,k}\end{pmatrix}-\boldsymbol{J}_k^{-1}\begin{pmatrix}\boldsymbol{J}_{01}\boldsymbol{\delta}\\ \pi_k\boldsymbol{J}_{11}\boldsymbol{\delta}\end{pmatrix}\\&=\boldsymbol{J}_k^{-1}\begin{pmatrix}\sqrt{n}\{\overline{U}_n-{\rm E}U_k(Y_1)\}\\ \sqrt{n}\{\overline{V}_{n,k}-{\rm E}V_k(Y_1)\}\end{pmatrix}.\end{aligned}$$

Now from [11, Lemmas 3.1 and 3.2], it follows that

$$\sqrt{n}(\overline{U}_n{\rm E}U_k(Y_1),\overline{V}_{n,k}-{\rm E}V_k(Y_1))\xrightarrow{\rm D}(M,N_k),$$

and the result holds.