The Annals of Statistics
2017, Vol. 45, No. 6, 2654–2679
https://doi.org/10.1214/17-AOS1538
© Institute of Mathematical Statistics, 2017

A WEIGHT-RELAXED MODEL AVERAGING APPROACH FOR HIGH-DIMENSIONAL GENERALIZED LINEAR MODELS

BY TOMOHIRO ANDO[*,1] AND KER-CHAU LI[†,‡,2]

University of Melbourne[*], *University of California, Los Angeles*[†] *and Academia Sinica*[‡]

Model averaging has long been proposed as a powerful alternative to model selection in regression analysis. However, how well it performs in high-dimensional regression is still poorly understood. Recently, Ando and Li [*J. Amer. Statist. Assoc.* **109** (2014) 254–265] introduced a new method of model averaging that allows the number of predictors to increase as the sample size increases. One notable feature of Ando and Li's method is the relaxation on the total model weights so that weak signals can be efficiently combined from high-dimensional linear models. It is natural to ask if Ando and Li's method and results can be extended to nonlinear models. Because all candidate models should be treated as working models, the existence of a theoretical target of the quasi maximum likelihood estimator under model misspecification needs to be established first. In this paper, we consider generalized linear models as our candidate models. We establish a general result to show the existence of pseudo-true regression parameters under model misspecification. We derive proper conditions for the leave-one-out cross-validation weight selection to achieve asymptotic optimality. Technically, the pseudo true target parameters between working models are not linearly linked. To overcome the encountered difficulties, we employ a novel strategy of decomposing and bounding the bias and variance terms in our proof. We conduct simulations to illustrate the merits of our model averaging procedure over several existing methods, including the lasso and group lasso methods, the Akaike and Bayesian information criterion model-averaging methods and some other state-of-the-art regularization methods.

1. Introduction.

In the process of statistical model building, researchers are often confronted with several candidate models to explore the data. Under the context of model selection, once one optimal model is chosen, the rest are discarded. However, if the aim of modeling is to predict future outcomes, then combining the predictions from different models by weighted averages increases prediction flexibility. One advantage of model averaging is its ability to incorporate model uncertainty. If the weights can be suitably determined, then better prediction may

Received December 2015; revised September 2016.

[1] Supported by Melbourne Business School, Australia.

[2] Supported in part by NSA Grant DMS-1513622 and in part by Academia Sinica, Taiwan.

MSC2010 subject classifications. Primary 62J12; secondary 62F99.

Key words and phrases. Asymptotic optimality, high-dimensional regression models, model averaging, model misspecification.

be obtained. Many studies on weight selection have been conducted under the classical setting, where the total number of model parameters is much smaller than the sample size, including those on the Akaike information criterion (AIC) [Akaike (1978, 1979)], Bayesian information criterion (BIC) model averaging [Min and Zellner (1992), Madigan and Raftery (1994), Kass and Raftery (1995), Raftery, Madigan and Hoeting (1997), Hoeting et al. (1999)], focused information criterion (FIC) model averaging [Claeskens and Hjort (2003), Hjort and Claeskens (2003)], Bayesian model averaging using predictive measures [Eklund and Karlsson (2007)], and predictive likelihood model averaging [Ando and Tsay (2010)]. Applications of the model averaging approach can be found in various fields; see, for example, Yeung, Bumgarner and Raftery (2005) in microarray data analysis, Montgomery and Nyhan (2010) in political science, Lee (2014) in operation management, Ando (2009), Ouysse and Kohn (2010) in asset pricing and Chung, Rust and Wedel (2009) in marketing.

Recent advances in information technology have greatly altered the data-accessing environment. It is now commonplace for statisticians to face high-dimensional data where the variables under study may outnumber the cases. Although intensive investigations have been conducted to address many challenges encountered, such studies are mostly concerned with model selection and shrinkage methods. Model averaging for high-dimensional regression is recently investigated by Ando and Li (2014).

One major feature of Ando and Li's (2014) model averaging approach allows the model weights to vary freely between 0 and 1 without the standard constraint of summing up to 1. In general, each candidate model may have its unique strength in capturing certain aspects of the signal. By integrating the information distilled from different models, the weight relaxation can substantially lower the smallest prediction errors achievable by model averaging. A theorem was established by Ando and Li to demonstrate the asymptotical optimality of weight selection by the leave-one-out cross-validation method. However, their paper was confined to linear regression models, thus limiting the application to the analysis of binary data, categorical or count variables as well as other practical situations.

In this paper, we study the model averaging approach for high-dimensional generalized linear models. The extension to generalized linear models involves several technical challenges to overcome. One difficulty concerns with the target parameters. For linear models, the target parameter exists for each candidate model and is clearly defined by linear projection. For generalized linear models, following Flynn, Hurvich and Simonoff (2013) and others, we call the theoretical target (namely, the parameter that minimizes the Kullback–Leibler loss between the specified model and the true data generating process) the pseudo true parameter. However, to the best of our knowledge, no studies have yet obtained conditions to guarantee the existence of the pseudo true parameter in each candidate model. To resolve the issue, we establish Theorem 1, which also provides the theoretical support to a number of earlier claims in studying generalized linear models under the context of model misspecification; for example, White (1982), Flynn, Hurvich and Simonoff (2013) and Lv and Liu (2014).

We use the Kullback–Leibler (KL) distance as a replacement for the squared prediction error to establish the asymptotic optimality of the leave-one-out cross-validation weight selection procedure. We notice that a straightforward extension of the strategy used in the proof of Ando and Li (2014) would lead to unsatisfactory results. In linear regression, the response variable can be directly decomposed into two parts, the true mean term and the random error term. The mean terms between the working models are neatly connected to each other via linear transformation of the true mean vector of the response. Each leave-one-out estimator is determined by a specific zero-diagonal matrix associated with the working model. However, such key properties used in Ando and Li (2014)'s derivation of the asymptotic optimality are no longer available in the generalized linear models. For example, the mean and variance of the response variable are both nonlinear functions of the regression parameters. Although local linearization can be performed via Taylor expansion, the pseudo true target parameters between working models are not linearly linked. Therefore, a different strategy for decomposing and bounding the bias and variance terms must be deployed to elevate the difficulties; see Remark 7 following the proof of Theorem 2.

We assess the performance of our new model-averaging procedure through simulation. The results indicate that our method yields more accurate predictions than many existing methods, including the lasso and group lasso methods, the AIC and BIC model-averaging methods and some other state-of-the-art methods. We also apply our proposed method to a Portuguese marketing campaign data set and demonstrate the good performance in prediction.

Recently, Charkhi, Claeskens and Hansen (2016) presented a different perspective of model averaging approach for general likelihood models, which complements our study. Their setting is built upon a framework of local model misspecification. The averaging criterion and the weight constraint are also different from ours.

The remainder of the paper is organized as follows. Section 2 considers the generalized linear model under model misspecification. We provide proper conditions for ensuring the existence of the pseudo true parameter. Section 3 introduces a new model-averaging procedure for high-dimensional generalized linear models, wherein the number of predictors may exceed the number of observations. Section 4 provides the theoretical results on the asymptotic optimality of weight selection by leave-one-out cross-validation. In Section 5, we present simulation results to illustrate the merit of the proposed method. The performance is compared with previously proposed model-averaging procedures, lasso and its variants. Additional discussion and the concluding remarks are provided in Section 6.

Notation. Let $\|A\| = [\mathrm{tr}\{A'A\}]^{1/2}$ be the norm of the matrix A, where "tr" denotes the trace of a square matrix. The big O and small o notation are used to indicate the order of a sequence relative to another sequence. For example, $a_n = O(b_n)$

states that the deterministic sequence a_n is at most of order b_n, while $c_n = o_p(d_n)$ states that c_n is a smaller order of d_n in probability.

2. Misspecified generalized linear models and the pseudo true parameter value.

We consider the one-parameter natural exponential family for constructing the working models

$$f(y|\theta) = \exp\{y\theta - b(\theta) + c(y)\}, \tag{2.1}$$

$$\Theta = \left\{\theta : \int \exp\{y\theta - b(\theta) + c(y)\}\, d\mu(y) < \infty\right\}. \tag{2.2}$$

Here, the measure $\mu(\cdot)$ can be continuous or discrete. By the convexity of $\exp(\cdot)$, Θ is a convex set. We assume that

(E1) Θ is an open set.

The open set Θ must be an open interval: $(-\infty, \infty)$, $(-\infty, \theta_1)$, (θ_1, ∞), or (θ_1, θ_2). Note that if $\theta_0 \in \Theta$, then we can re-parameterize θ by taking the difference $\theta - \theta_0$ and adjusting $c(y)$ accordingly. Later on, without loss of generality, we may assume that

(E1)* Θ is an open set which contains 0.

It is well known that $E[Y|\theta] = b'(\theta)$ and $\mathrm{Var}[Y|\theta] = b''(\theta)$ if the true distribution of Y belongs to the natural exponential family (2.1). However, the true density of Y may not follow (2.1). Studies on the quasi maximum likelihood estimator have been mostly based on asymptotic results, including consistency and asymptotic normality, among others [see White (1982)]. The issue of how to ensure the existence of the pseudo true parameter of the quasi maximum likelihood estimator has not been addressed.

Suppose we have n observations $\{(y_i, \mathbf{x}_i); i = 1, 2, \dots, n\}$, where y_i is the response variable and $\mathbf{x}_i$ is a p-dimensional vector of explanatory variables. To construct a working model, the natural parameter in (2.1) is linked to the predictors via the equation

$$\theta = \boldsymbol{\beta}'\mathbf{x}, \tag{2.3}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector. Denote $b'(\Theta) = \{b'(\theta) : \theta \in \Theta\}$. The following assumptions will be made:

(E2) $y_1, \dots, y_n$ are independent random variables with means $\mu_i = E[Y_i] \in b'(\Theta)$.

(E3) $\mathbf{x}_1, \dots, \mathbf{x}_n \in R^p$ are fixed. The dimension of the space spanned by $\mathbf{x}_i$ is p.

(E4) The set $B = \{\boldsymbol{\beta} : \boldsymbol{\beta}'\mathbf{x}_i \in \Theta, i = 1, \dots, n\}$ is nonempty.

The following theorem establishes the existence of the pseudo-true regression parameter in a working model.

THEOREM 1. *Under* (E1)∼(E4),

$$\sup_{\boldsymbol{\beta}\in B}\left[\sum_{i=1}^{n}\mu_i\mathbf{x}_i'\boldsymbol{\beta}-\sum_{i=1}^{n}b(\mathbf{x}_i'\boldsymbol{\beta})\right] \tag{2.4}$$

is finite and is achieved uniquely at some point $\boldsymbol{\beta}=\boldsymbol{\beta}^*$.

Following from Theorem 1, the pseudo-true parameter must satisfy the likelihood equation.

COROLLARY. *Under* (E1)∼(E4), *the equation*

$$\sum_{i=1}^{n}\mu_i\mathbf{x}_i-\sum_{i=1}^{n}b'(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i=\mathbf{0} \tag{2.5}$$

has a unique solution $\boldsymbol{\beta}=\boldsymbol{\beta}^*\in B$.

PROOF OF THEOREM 1. To prove this theorem, the following strategy is used. Consider a concave function $G(\cdot)$ defined on an open set $B\in R^p$. Take a point $\boldsymbol{\beta}_0$ in B, and consider any vector $\mathbf{e}$ with unit length $\|\mathbf{e}\|=1$. Let U_e be the set of c such that $c\mathbf{e}\in B$. It is obvious that

$$\sup_{\boldsymbol{\beta}\in B}G(\boldsymbol{\beta})=\sup_{\mathbf{e}}\sup_{c\in U_e}G(\boldsymbol{\beta}_0+c\mathbf{e}). \tag{2.6}$$

To show that the maximum is achieved at a finite point $\boldsymbol{\beta}=\boldsymbol{\beta}^*$, it suffices to show that

(E5) $\sup_{c\in U_e}G(\boldsymbol{\beta}_0+c\mathbf{e})$ is finite and achieved at a unique value $c=c(\mathbf{e})$.
(E6) $c(\mathbf{e})$ is continuous in $\mathbf{e}$.

Following from Lemmas 1 and 2 given below, we have completed the proof of Theorem 1. □

LEMMA 1. *For the concave function* $G(\boldsymbol{\beta})=\sum_{i=1}^{n}\mu_i\mathbf{x}_i'\boldsymbol{\beta}-\sum_{i=1}^{n}b(\mathbf{x}_i'\boldsymbol{\beta})$, (E5) *holds.*

LEMMA 2. *For any concave function* $G(\cdot)$, *if* (E5) *holds*, *then* (E6) *holds.*

The proofs of Lemmas 1 and 2 are given in the Appendix.

REMARK 1. In the literature [e.g., Flynn, Hurvich and Simonoff (2013), Lv and Liu (2014)], the pseudo true parameter is simply defined as the solution to the score equation (2.5). However, the existence of pseudo true parameter is not guaranteed. For example, if the working model is specified by the exponential distribution, then our theorem establishes existence only when the true mean μ_i

$(i = 1, \dots, n)$ are positive. Thus, if we encounter a situation where the response variables are likely to have negative values, then condition (E2) may not hold and we should avoid the use of exponential distributions for constructing candidate models.

In the context of model selection for nonlinear models, it is often necessary to assume the existence of quasi maximum likelihood estimators for all sub-models if the true distribution is contained in a complete model. Clearly, this assumption cannot hold without suitable conditions. For generalized linear models, our Theorem 1 characterizes a minimum set of conditions required to justify the existence of quasi MLE.

3. A new model-averaging procedure. There are two steps involved in our model-averaging procedure.

3.1. *Step* 1: *Preparation of the candidate models.* We denote a set of M candidate models $M_1, \dots, M_M$ by

$$M_k : f(y|\boldsymbol{\beta}_k, \mathbf{x}_k) = \exp\{y(\boldsymbol{\beta}_k'\mathbf{x}_k) - b(\boldsymbol{\beta}_k'\mathbf{x}_k) + c(y)\}, \tag{3.1}$$

where $\boldsymbol{\beta}_k$ is the parameter vector for the p_k-dimensional predictor $\mathbf{x}_k$ of model M_k, $k = 1, \dots, M$. Under the candidate model M_k, the maximum likelihood estimate of the regression coefficients $\boldsymbol{\beta}_k$ is given by

$$\hat{\boldsymbol{\beta}}_k = \operatorname*{argmax}_{\beta_k} \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}_k, \mathbf{x}_{ki}),$$

wherein $\mathbf{x}_{ki}$ denotes the predictors associated with outcome y_i under model M_k. This yields the MLE estimates, $\hat{\eta}_{ki} = \mathbf{x}_{ki}'\hat{\boldsymbol{\beta}}_k$ for $i = 1, \dots, n$, which can be used to predict $E(y_i)$ via the link $b'(\hat{\eta}_{ki})$ under model M_k.

After a set of M candidate models is specified and their maximum likelihood estimates $\{\hat{\boldsymbol{\beta}}_1'\mathbf{x}_1, \dots, \hat{\boldsymbol{\beta}}_M'\mathbf{x}_M\}$ are obtained, we need to determine the weight of each model. The determination of the weight is important in model averaging because it directly affects the performance of the model-averaging estimator. We allow the M-dimensional weight vector $\mathbf{w} = (w_1, \dots, w_M)'$ to be chosen from the unit hypercube of R^M:

$$Q_n = \{\mathbf{w} \in [0, 1]^M : 0 \le w_k \le 1\}.$$

For $i = 1, \dots, n$, our model-averaging estimates can be expressed by

$$\eta(\mathbf{x}_i; \mathbf{w}, \hat{\boldsymbol{\beta}}) = \sum_{k=1}^{M} w_k \hat{\boldsymbol{\beta}}_k' \mathbf{x}_{ki}, \tag{3.2}$$

where $\hat{\boldsymbol{\beta}}' = (\hat{\boldsymbol{\beta}}_1', \dots, \hat{\boldsymbol{\beta}}_M')$. Using the model-averaging estimates, we can predict $E(y_i)$ by $b'(\eta(\mathbf{x}_i; \mathbf{w}, \hat{\boldsymbol{\beta}}))$.

3.2. *Step* 2: *Optimal weight selection by leave-one-out cross-validation.* For $k = 1, \dots, M$, let $\hat{\boldsymbol{\beta}}_{k(-i)}$ denote the jackknife estimator of $\boldsymbol{\beta}_k$ for model M_k with the ith observation deleted. We define the leave-one-out cross-validation criterion function

$$CV(\mathbf{w}) = \sum_{i=1}^{n} \left[y_i \left(\sum_{k=1}^{M} w_k \hat{\boldsymbol{\beta}}'_{k(-i)} \mathbf{x}_{ki} \right) - b \left(\sum_{k=1}^{M} w_k \hat{\boldsymbol{\beta}}'_{k(-i)} \mathbf{x}_{ki} \right) + c(y_i) \right] \tag{3.3}$$

and choose the weights by maximizing the objective function $CV(\mathbf{w})$ over the set Q_n. Because $CV(\mathbf{w})$ is convex in $\mathbf{w}$, the global optimization can be performed efficiently through constrained optimization programming. For example, the optim package in the R language and other open software packages can be applied for this purpose.

Let $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_M)'$ be the maximizer of $CV(\mathbf{w})$. We use $\sum_{k=1}^{M} \hat{w}_k \hat{\boldsymbol{\beta}}'_k \mathbf{x}_{ki}$ as our final model-averaging estimate of the natural parameter for y_i, yielding the final model

$$\begin{aligned} & f(y_i | \hat{\mathbf{w}}, \hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_M) \\ & \quad = \exp\left\{ y_i \left(\sum_{k=1}^{M} \hat{w}_k \hat{\boldsymbol{\beta}}'_k \mathbf{x}_{ki} \right) - b \left(\sum_{k=1}^{M} \hat{w}_k \hat{\boldsymbol{\beta}}'_k \mathbf{x}_{ki} \right) + c(y_i) \right\}. \end{aligned} \tag{3.4}$$

REMARK 2. To prepare a set of candidate models without prior subject knowledge or expert theory, we can employ the grouping procedure based on marginal information [Ando and Li (2014)]. We calculate the p-values of each predictor x_j ($j = 1, \dots, p$) by fitting the generalized linear model with a single predictor by setting the natural parameter in (2.3) as $\theta = \alpha_j + \beta_j x_j$. Sorting the set of p predictors based on the p-values, we partition the set of p predictors into $M + 1$ groups. The first group has the lowest p-values, and the $(M + 1)$th group has the values closest to one. Each group is used for constructing a candidate model, yielding models M_i, for $i = 1, \dots, M$. The $(M + 1)$th group, which is expected to contain only insignificant variables, is discarded.

REMARK 3. Because of the ordering of predictors by marginal information, the first few models are likely to be more informative than the last few models. If we were to impose $\sum_{k=1}^{M} w_k = 1$, then some weights may be shifted away from the first few models, which may lead to substantial bias. Ando and Li (2014) provided an example, where the optimal weight assignment should be $\mathbf{w} = (1, 1, \dots, 1)'$. Further discussion is given later in Section 4.4.

4. Theoretical results for the model averaging estimator. This section investigates some properties of the proposed model-averaging procedure. After recasting the model-averaging problem, we define the pseudo true regression parameters for candidate models using the result of Theorem 1. Then we describe the set of assumptions needed for establishing the asymptotic optimality of the weight selection procedure by the leave-one-out cross-validation. The main theorem on optimality is given in Section 4.3. Section 4.4 provides further discussion on the impact of weight constraint and the class structure of candidate models. A study on optimal grouping of the predictors for risk minimization is also given in the supplementary document [Ando and Li (2017)].

4.1. *The structure of pseudo true regression parameters under model specification.* Recall that all predictors are considered as fixed in this paper. Given the candidate model M_k, we fit the output $y_i, i = 1, \dots, n$ by the maximizing the joint likelihood $\prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}_k, \mathbf{x}_{ki})$. Because we treat candidate models only as working models, the lack of model fit may come from two distinct sources of model misspecification. The first source is the discrepancy between the true density of y_i and the proposed one-parameter exponential family (2.1). Let η_i be the pseudo true parameter of θ for fitting y_i with (2.1), namely

$$\eta_i = \underset{\theta}{\operatorname{argmax}}\, E \log f(y_i|\theta) = \underset{\theta}{\operatorname{argmax}}\, E\{y_i\theta - b(\theta)\}. \tag{4.1}$$

The assumption (E2) in Section 2 guarantees the existence of η_i. The second source of model misspecification lies in approximating η_i by the canonical link (2.3), wherein p is replaced by p_k, the number of variables in model M_k regression parameter $\boldsymbol{\beta}_k'\mathbf{x}_{ki}$. Under the assumptions of (E1)$\sim$(E4) (with p being replaced by p_k), we can use Theorem 1 to establish the existence of the pseudo-true regression parameter $\boldsymbol{\beta}_{k0}$ [namely the solution of the equation (2.5) given in the Corollary].

4.2. *Assumptions for asymptotic study of model averaging.* In this section, we put together a set of assumptions for investigating asymptotic behavior of the proposed weight selection procedure for model averaging.

Assumptions on candidate models: (R1) Conditions (E1)* and (E2) hold. Conditions (E3) $\sim$(E4) (with p being replaced by p_k) hold for each candidate model $M_k, k = 1, \dots, M_M$.

(R2) The second derivative of $b(\boldsymbol{\beta}_{k0}'\mathbf{x}_{ki})$ is continuous. Additionally, for some constants $C_1, C_2 > 0$ that do not depend on n,

$$\sup_{k,n}\left| n^{-1}\sum_{i=1}^{n}(b'''(\boldsymbol{\beta}_{k0}'\mathbf{x}_{ki}))\right| < C_2 < \infty, \tag{4.2}$$

$$0 < C_1 < \inf_{k,n,\mathbf{w}\in Q_n}\left\{ b''\left(\sum_{k=1}^{M} w_k \boldsymbol{\beta}_{k0}'\mathbf{x}_{ki}\right)\right\}, \qquad i = 1, \dots, n, \tag{4.3}$$

$$\sup_{k,n,\mathbf{w}\in Q_n}\left(n^{-1}\sum_{i=1}^{n} b''\left(\sum_{k=1}^{M} w_k \boldsymbol{\beta}_{k0}'\mathbf{x}_{ki}\right)\right) < C_2 < \infty \tag{4.4}$$

hold. Here, recall that $\boldsymbol{\beta}_{k0}$ is the pseudo-true regression parameter that minimizes the KL distance measure between the true model and the working model M_k.

REMARK 4. Assumption (R1) is the minimal requirement for rationalizing which exponential family to employ. Conditions (E1)* and (E2) are elementary and they should be justified beforehand. On the other hand, if conditions (E3)∼(E4) do not hold for a particular candidate model, then such a model should be eliminated automatically from the candidate list. We note that these conditions are also critical for studying traditional model selection procedures under model-misspecification.

Assumptions concerning KL distance risk Given a weight vector $\mathbf{w}$, we measure the divergence of using the model average estimate, $\sum_{k=1}^{M} w_k \hat{\boldsymbol{\beta}}_k' \mathbf{x}_{ki}$ to approximate η_i, $i = 1, \dots, n$ by considering the KL distance. We define

$$
\begin{aligned}
L(\mathbf{w}) = \sum_{i=1}^{n} \Biggl\{ & b'(\eta_i) \Biggl[\eta_i - \sum_{k=1}^{M} w_k \hat{\boldsymbol{\beta}}_k' \mathbf{x}_{ki} \Biggr] \\
& - \Biggl[b(\eta_i) - b\Biggl(\sum_{k=1}^{M} w_k \hat{\boldsymbol{\beta}}_k' \mathbf{x}_{ki} \Biggr) \Biggr] \Biggr\},
\end{aligned} \tag{4.5}
$$

and the expected KL distance

$$R(\mathbf{w}) = E_Y L(\mathbf{w}),$$

where the expectation is calculated with respect to the joint density of $(y_1, \dots, y_n)$ conditional on the predictors.

Define

$$\varepsilon_i = y_i - \mu_i = y_i - E(y_i)$$

and

$$\zeta_n = \inf_{\mathbf{w} \in Q_n} R(\mathbf{w}),$$

the infeasible minimum value of $R(\mathbf{w})$.

The following additional assumptions will be needed in proving Theorem 2. For some fixed integer $1 \le K < \infty$,

$$\text{(A1)} \quad E[\varepsilon_i^{4K}] \le B < \infty, \qquad i = 1, \dots, n,$$

$$\text{(A2)} \quad \sup_k \frac{1}{p_k} \bar{\lambda}\{H_k(\boldsymbol{\beta}_{k0})\} \le \Gamma n^{-1},$$

$$\text{(A3)} \quad M^{4K+2} n^K / \zeta_n^{2K} \to 0,$$

$$\text{(A4)} \quad 0 < C_1 < \frac{1}{n} \sum_{i=1}^{n} \mu_i^2 < C_2 < \infty,$$

$$\text{(A5)} \quad \sup_{1 \le k \le M} \frac{p_k}{M^{1+1/K} n^{1/2}} \le \Lambda < \infty.$$

Here, B, Γ, Λ, C_1 and C_2 are some constants, and $\bar{\lambda}\{\cdot\}$ denotes the maximal diagonal element of a matrix, $H_k(\boldsymbol{\beta}_k) = B_k^{1/2} X_k (X_k' B_k X_k)^{-1} X_k' B_k^{1/2}$, where B_k is an $n \times n$ diagonal matrix with ith element $B_{k,ii} = b''(\mathbf{x}_{ki}' \boldsymbol{\beta}_k)$.

REMARK 5. We clarify the relationship between the set of assumptions (A1)–(A5) and the set of conditions (4)–(8) made in Ando and Li (2014), for linear regression. The moment condition (A1) corresponds to condition (4) in Ando and Li (2014). Under the linear regression case, the $p_k \times p_k$ matrix $H_k(\boldsymbol{\beta}_k)$ is given as $H_k(\boldsymbol{\beta}_k) = X_k(X_k' X_k)^{-1} X_k'$. Thus, the condition (A2) corresponds to the condition (5) of Ando and Li. Condition (A3) and (A4) are the same as the conditions (7) and (8) in Ando and Li. The only condition which is stronger than Ando and Li is the condition (A5). To compare (A5) with the condition (6) in Ando and Li, we use (A3) to replace M by $M = o(\zeta_n^{K/(2K+1)}/n^{K/(4K+2)})$ in (A5), obtaining

$$\sup_{1 \le k \le M} p_k / \big[n^{1/2 - (\frac{1}{4+2/K})(1+\frac{1}{K})} \zeta_n^{(\frac{1}{2+1/K})(1+\frac{1}{K})}\big] = o(1).$$

Suppose the error ε_i is well behaved and the moment condition (A1) holds for a very large K. By taking $1/K \approx 0$, the above expression becomes

$$\sup_{1 \le k \le M} \frac{p_k}{n^{1/4} \zeta_n^{1/2}} = o(1).$$

As assumed in Ando and Li (2014), if we assume ζ_n converges to ∞ at the rate of $n^{1-\delta}$ for $0 < \delta < 1/2$, then the condition (A5) is simplified to $\sup_{1 \le k \le M} p_k / n^{3/4-\delta/2} = o(1)$, which is slightly stronger than condition (6) of Ando and Li; $\sup_{1 \le k \le M} p_k / n^{3/4} \le \Lambda < \infty$. It is also noted that (A3) is reduced to $M = o(n^{1/4-\delta/2})$ which gives a bound on the diverging order of M.

Theorem 2 focuses on the asymptotic optimality of $\mathbf{w}$ after choosing the grouping of the predictors by marginal screening. Lemma 4 (in the supplementary document) shows that partitioning after ordering the likelihood is an optimal way of generating models for model averaging.

4.3. *Asymptotic optimality of the weight selection procedure.* Following Li (1987), Hansen (2007), Wan, Zhang and Zou (2010), Hansen and Racine (2012) and Ando and Li (2014), we demonstrate that our weight selection procedure is asymptotically optimal in certain sense. Recently, Zhang, Li and Tsai (2010) and Flynn, Hurvich and Simonoff (2013) considered the asymptotic loss efficiency of shrinkage methods in the context of generalized linear models and adopted the KL distance measure to define the risk of a parameter estimate. For our model-averaging procedure, the KL distance measure between the true model and the estimated model $f(y|\hat{\mathbf{w}}, \hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_M, \mathbf{x})$ in (3.4) is given in (4.5). Ideally, if the true $\eta_i (i = 1, \dots, n)$ is known, we can choose $\mathbf{w}$ to minimize $L(\mathbf{w})$.

Theorem 2 shows that the cross-validation $CV(\mathbf{w})$ almost behaves like a procedure that provides the lowest $L(\mathbf{w})$ among all weight choices. Similar to Li (1986, 1987), we show that the smallest possible KL distance $\inf_{\mathbf{w}\in Q_n} L(\mathbf{w})$, which is infeasible to achieve because the true distribution is unknown, is achievable by cross-validation. Ando and Li (2014) demonstrated this property in the context of high-dimensional linear regression models. Thus, our result is a natural extension of Theorem 1 in Ando and Li (2014) to averaging in the exponential family. However, great difficulties were encountered when applying the strategy used in the proof by Ando and Li (2014).

THEOREM 2. *Assume that the regularity conditions* (R1)–(R2) *and* (A1)–(A5) *hold. Then*, $\hat{\mathbf{w}}$ *is asymptotically optimal in the sense that*

$$\frac{L(\hat{\mathbf{w}})}{\inf_{\mathbf{w}\in Q_n} L(\mathbf{w})} \to 1,$$

where the convergence is in probability.

REMARK 6. We can also obtain the asymptotic optimality for the model averaging estimator under the restriction $H_n = \{0 \le w_k; \sum_{k=1}^{M} w_k = 1\}$. Because the weight restriction H_n is stronger than Q_n, the assumption (A3) can be weakened. For more details, see Appendix E in the supplementary document.

To prove Theorem 2, we need the following lemma. This lemma considers the stochastic relationship between $\tilde{\boldsymbol{\eta}}_k = (\mathbf{x}'_{k1}\hat{\boldsymbol{\beta}}_{k(-1)}, \mathbf{x}'_{k2}\hat{\boldsymbol{\beta}}_{k(-2)}, \dots, \mathbf{x}'_{kn}\hat{\boldsymbol{\beta}}_{k(-n)})'$ and $\hat{\boldsymbol{\eta}}_k = (\mathbf{x}'_{k1}\hat{\boldsymbol{\beta}}_k, \mathbf{x}'_{k2}\hat{\boldsymbol{\beta}}_k, \dots, \mathbf{x}'_{kn}\hat{\boldsymbol{\beta}}_k)$. The proof of Lemma 3 is given in the supplementary document.

LEMMA 3. *Assume that* (R1), (R2), (A2), (A3) *and* (A5) *hold. Then the relation between* $\tilde{\boldsymbol{\eta}}_k$ *and* $\hat{\boldsymbol{\eta}}_k$ *is*

$$\tilde{\boldsymbol{\eta}}_k - \hat{\boldsymbol{\eta}}_k = S_k\hat{\mathbf{v}}_k + o_p(1), \tag{4.6}$$

where S_k *is a diagonal matrix with the* i*th element equal to* $h_{k,ii}/(1-h_{k,ii})$ *and* $\hat{\mathbf{v}}_k$ *is an* n*-dimensional vector with* i*th element* $\{y_i - b'(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_k)\}/b''(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_k)$.

PROOF OF THEOREM 2. We will share the bounding constants (C, C', etc.) when deriving inequalities. From Assumptions (R1), for each of the individual models $M_1, \dots, M_M$, we can ensure a unique minimizer of the KL distance measure between the true model and the estimated kth individual model in (3.1). Let $\boldsymbol{\beta}_{k0}$ be the unique minimizer; then, it satisfies $X'_k(\boldsymbol{\mu} - \mathbf{b}'(X_k\boldsymbol{\beta}_{k0})) = \mathbf{0}$ for $k = 1, \dots, M$. Here, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ is the true mean vector, and $\mathbf{b}'(X_k\boldsymbol{\beta}_{k0}) =$

$(b'(\mathbf{x}'_{k1}\boldsymbol{\beta}_{k0}),\dots,(b'(\mathbf{x}'_{kn}\boldsymbol{\beta}_{k0})))'$. The following notation will be used:

$$\begin{aligned}\hat{\boldsymbol{\eta}}_k &= (\hat{\eta}_{k1},\dots,\hat{\eta}_{kn})' = (\hat{\boldsymbol{\beta}}'_k\mathbf{x}_{k1},\dots,\hat{\boldsymbol{\beta}}'_k\mathbf{x}_{kn})',\\ \tilde{\boldsymbol{\eta}}_k &= (\tilde{\eta}_{k1},\dots,\tilde{\eta}_{kn})' = (\hat{\boldsymbol{\beta}}'_{k(-1)}\mathbf{x}_{k1},\dots,\hat{\boldsymbol{\beta}}'_{k(-n)}\mathbf{x}_{kn})',\\ \boldsymbol{\eta}_{k0} &= (\eta_{k01},\dots,\eta_{k0n})' = (\boldsymbol{\beta}'_{k0}\mathbf{x}_{k1},\dots,\boldsymbol{\beta}'_{k0}\mathbf{x}_{kn})'.\end{aligned}$$

We begin with the connection between $CV(\mathbf{w})$ and $L(\mathbf{w})$:

$$-CV(\mathbf{w}) = \tilde{L}(\mathbf{w}) - B(\mathbf{w}) - \sum_{i=1}^{n}\{b'(\eta_i)\eta_i + \varepsilon_i\eta_i - b(\eta_i) - c(y_i)\},$$

where

$$\begin{aligned}\tilde{L}(\mathbf{w}) &= \sum_{i=1}^{n}\left\{b'(\eta_i)\left[\eta_i - \sum_{k=1}^{M}w_k\tilde{\eta}_{ki}\right] - \left[b(\eta_i) - b\left(\sum_{k=1}^{M}w_k\tilde{\eta}_{ki}\right)\right]\right\},\\ B(\mathbf{w}) &= \sum_{i=1}^{n}\left\{\varepsilon_i\left[\sum_{k=1}^{M}w_k\tilde{\eta}_{ki} - \eta_i\right]\right\}.\end{aligned}$$

Because $\hat{\mathbf{w}}$ minimizes $-CV(\mathbf{w})$ over $\mathbf{w} \in Q_n$, it also minimizes $\tilde{L}(\mathbf{w}) - B(\mathbf{w})$ (because the other terms are unrelated to $\mathbf{w}$). The claim $L(\hat{\mathbf{w}})/\inf_{\mathbf{w}\in Q_n} L(\mathbf{w}) \to 1$ is valid if

$$\sup_{\mathbf{w}\in Q_n}\left|\tilde{L}(\mathbf{w})/L(\mathbf{w}) - 1\right| \to 0, \tag{4.7}$$

$$\sup_{\mathbf{w}\in Q_n}\left|B(\mathbf{w})/R(\mathbf{w})\right| \to 0, \tag{4.8}$$

$$\sup_{\mathbf{w}\in Q_n}\left|L(\mathbf{w})/R(\mathbf{w}) - 1\right| \to 0 \tag{4.9}$$

hold. From these claims, we can see that the cross-validation criterion $CV(\mathbf{w})$ yields an unbiased estimate of the risk $R(\mathbf{w})$ up to a constant term independent of $\mathbf{w}$ asymptotically.

We first prove claim (4.7). We have

$$\begin{aligned}&\left|\tilde{L}(\mathbf{w}) - L(\mathbf{w})\right|\\ &\quad= \left|\sum_{i=1}^{n}b'(\eta_i)\left(\sum_{k=1}^{M}w_k\hat{\eta}_{ki} - \sum_{k=1}^{M}w_k\tilde{\eta}_{ki}\right)\right.\\ &\qquad\left. - \sum_{i=1}^{n}\left[b\left(\sum_{k=1}^{M}w_k\hat{\eta}_{ki}\right) - b\left(\sum_{k=1}^{M}w_k\tilde{\eta}_{ki}\right)\right]\right|\\ &\quad\le \left|\sum_{i=1}^{n}b'(\eta_i)\left(\sum_{k=1}^{M}w_k\hat{\eta}_{ki} - \sum_{k=1}^{M}w_k\tilde{\eta}_{ki}\right)\right|\end{aligned}$$

$$+\left|\sum_{i=1}^{n}\left[b\left(\sum_{k=1}^{M} w_k\hat{\eta}_{ki}\right)-b\left(\sum_{k=1}^{M} w_k\tilde{\eta}_{ki}\right)\right]\right|$$

$$=I_{11}(\mathbf{w})+I_{12}(\mathbf{w}).$$

By the Cauchy–Schwarz inequality,

$$I_{11}(\mathbf{w})\le\left(\sum_{i=1}^{n}b'(\eta_i)^2\right)^{1/2}\times\left(\sum_{i=1}^{n}\left[\sum_{k=1}^{M}w_k(\hat{\eta}_{ki}-\tilde{\eta}_{ki})\right]^2\right)^{1/2}.$$

To obtain the claim (4.7), we therefore need to prove

$$n^{1/2}\times\sup_{\mathbf{w}\in Q_n}\left[\sum_{i=1}^{n}\left|\sum_{k=1}^{M}w_k(\hat{\eta}_{ki}-\tilde{\eta}_{ki})\right|^2\right]^{1/2}\Big/L(\mathbf{w})\to 0. \tag{4.10}$$

We can bound $\sum_{i=1}^{n}|\sum_{k=1}^{M}w_k(\hat{\eta}_{ki}-\tilde{\eta}_{ki})|^2$ by

$$M^2\max_{k=1,\dots,M}\sum_{i=1}^{n}(\hat{\eta}_{ki}-\tilde{\eta}_{ki})^2.$$

Here, we used the triangle inequality to first get an upper bound, $(\sum_{k=1}^{M}w_k\|\hat{\boldsymbol{\eta}}_k-\tilde{\boldsymbol{\eta}}_k\|)^2\le M^2\max_{k=1,\dots,M}\|\hat{\boldsymbol{\eta}}_k-\tilde{\boldsymbol{\eta}}_k\|^2$. Therefore, using (4.9), it is sufficient to show

$$n^{1/2}\times\left(M^2\max_{k=1,\dots,M}\sum_{i=1}^{n}(\hat{\eta}_{ki}-\tilde{\eta}_{ki})^2\right)^{1/2}\Big/\zeta_n\to 0. \tag{4.11}$$

Using (4.6) in Lemma 3, we have $\tilde{\boldsymbol{\eta}}_k-\hat{\boldsymbol{\eta}}_k=S_k\hat{\mathbf{v}}_k+o_p(p_k^{1/2}/n^{1/2})$, where S_k is a diagonal matrix with the ith element equal to $h_{k,ii}/(1-h_{k,ii})$, $h_{k,ii}$ is the ith diagonal element of the working matrix $H_k=B_k^{1/2}X_k(X_k'B_kX_k)^{-1}X_k'B_k^{1/2}$, and $\hat{\mathbf{v}}_k$ is an n-dimensional vector with ith element $\hat{v}_{ki}=\{y_i-b'(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_k)\}/b''(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_k)$. Thus, for some positive constant C,

$$\|\tilde{\boldsymbol{\eta}}_k-\hat{\boldsymbol{\eta}}_k\|^2\le C\times\{\lambda_{\max}(S_k)\}^2\|\hat{\mathbf{v}}_k\|^2\le O_p\left\{\left(\frac{\bar{\lambda}(H_k)}{1-\bar{\lambda}(H_k)}\right)^2\times n\right\}=O_p\left(\frac{p_k^2}{n}\right),$$

where $\bar{\lambda}(H_k)$ is the maximal diagonal element of H_k. Thus,

$$n^{1/2}\times\left(M^2\max_{k=1,\dots,M}\sum_{i=1}^{n}(\hat{\eta}_{ki}-\tilde{\eta}_{ki})^2\right)^{1/2}\Big/\zeta_n\le C'\times\frac{M\sup_k p_k}{\zeta_n}\to 0,$$

for some positive constant C'. Here, the last line follows from conditions (A3) and (A5). Therefore, $\sup_{\mathbf{w}\in Q_n}I_{11}(\mathbf{w})/L(\mathbf{w})\to 0$ is obtained.

We next prove $\sup_{\mathbf{w}\in Q_n}I_{12}(\mathbf{w})/L(\mathbf{w})\to 0$. From the Taylor expansion, $b(\sum_{k=1}^{M}w_k\tilde{\eta}_{ki})\approx b(\sum_{k=1}^{M}w_k\hat{\eta}_{ki})+b'(\sum_{k=1}^{M}w_k\hat{\eta}_{ki})(\sum_{k=1}^{M}w_k\hat{\eta}_{ki}-$

$\sum_{k=1}^M w_k\tilde{\eta}_{ki}) + \frac{1}{2}b''(\sum_{k=1}^M w_k\hat{\eta}_{ki})(\sum_{k=1}^M w_k\tilde{\eta}_{ki} - \sum_{k=1}^M w_k\hat{\eta}_{ki})^2$, it is enough to show that

$$\sup_{\mathbf{w}\in Q_n}\left|\sum_{i=1}^n b'\left(\sum_{k=1}^M w_k\hat{\eta}_{ki}\right)\left(\sum_{k=1}^M w_k\hat{\eta}_{ki} - \sum_{k=1}^M w_k\tilde{\eta}_{ki}\right)\Big/ R(\mathbf{w})\right| \to 0 \tag{4.12}$$

and

$$\sup_{\mathbf{w}\in Q_n}\left|\sum_{i=1}^n b''\left(\sum_{k=1}^M w_k\hat{\eta}_{ki}\right)\left(\sum_{k=1}^M w_k\tilde{\eta}_{ki} - \sum_{k=1}^M w_k\hat{\eta}_{ki}\right)^2\Big/ R(\mathbf{w})\right| \to 0. \tag{4.13}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\sum_{i=1}^n b'\left(\sum_{k=1}^M w_k\hat{\eta}_{ki}\right)\left(\sum_{k=1}^M w_k\hat{\eta}_{ki} - \sum_{k=1}^M w_k\tilde{\eta}_{ki}\right)\Big/ R(\mathbf{w}) \\
&\quad \le \left(\sum_{i=1}^n b'\left(\sum_{k=1}^M w_k\hat{\eta}_{ki}\right)^2\right)^{1/2} \times \left(\sum_{i=1}^n\left[\sum_{k=1}^M w_k(\hat{\eta}_{ki} - \tilde{\eta}_{ki})\right]^2\right)^{1/2}\Big/ R(\mathbf{w}) \\
&\quad \le C \times n^{1/2} \times \left(\frac{M^2 \sup_k p_k^2}{n}\right)^{1/2} / \zeta_n \to 0,
\end{aligned}$$

which is the claim (4.12). From (4.4), we can derive

$$\begin{aligned}
C_1 \times &\left(\sum_{i=1}^n\left[\sum_{k=1}^M w_k(\hat{\eta}_{ki} - \tilde{\eta}_{ki})\right]^2\right) \\
&\le \left(\sum_{i=1}^n b''\left(\sum_{k=1}^M w_k\hat{\eta}_{ki}\right)\left(\sum_{k=1}^M w_k\tilde{\eta}_{ki} - \sum_{k=1}^M w_k\hat{\eta}_{ki}\right)^2\right) \\
&\le C_2 \times \left(\sum_{i=1}^n\left[\sum_{k=1}^M w_k(\hat{\eta}_{ki} - \tilde{\eta}_{ki})\right]^2\right),
\end{aligned}$$

where C_1 and C_2 are some constants. Therefore, we can obtain the claim (4.13) by

$$\begin{aligned}
&\sup_{\mathbf{w}\in Q_n}\left|\sum_{i=1}^n b''\left(\sum_{k=1}^M w_k\hat{\eta}_{ki}\right)\left(\sum_{k=1}^M w_k\tilde{\eta}_{ki} - \sum_{k=1}^M w_k\hat{\eta}_{ki}\right)^2\Big/ R(\mathbf{w})\right| \\
&\quad \le C \times \sup_{\mathbf{w}\in Q_n}\left|\sum_{i=1}^n\left[\sum_{k=1}^M w_k(\hat{\eta}_{ki} - \tilde{\eta}_{ki})\right]^2\Big/ R(\mathbf{w})\right| \\
&\quad \le C \times \left(\frac{M^2 \sup_k p_k^2}{n}\right) \times \left(\frac{1}{\zeta_n}\right) \to 0,
\end{aligned}$$

where the last line follows from conditions (A3) and (A5). Therefore, we proved the claim $\sup_{\mathbf{w}\in Q_n} I_{12}(\mathbf{w})/R(\mathbf{w}) \to 0$. The claim (4.7) is obtained. The claims (4.8) and (4.9) are proved in the online supplementary document. This completes the proof of Theorem 2. □

REMARK 7. Because of the nonlinearity of $b'(\theta)$, we have employed a different strategy for error term decomposition in many places in Theorem 2. For example, the method of bounding the term $|\tilde{L}(\mathbf{w}) - L(\mathbf{w})|$, a critical step in the proof of (12) in Ando and Li (2014), is not applicable here. To overcome this hurdle, we break it into two terms $I_{11}(\mathbf{w})$ and $I_{12}(\mathbf{w})$ and then apply Lemma 3.

4.4. *Optimal weights under weight relaxation.* With the relaxation of the weights from adding up to 1, what do the optimal weights look like? Will they sum up to 1 asymptotically? To shed some light on these difficult questions raised by the Associate Editor, we consider two different scenarios of preparing candidate linear models for averaging. Denote the projection matrix $X_k(X_k'X_k)^{-1}X_k'$ associated with model M_k by H_k and write the weighted matrix $H(\mathbf{w}) = \sum_{k=1}^{M} w_k H_k$. The first scenario considers the extreme situation where the projection matrices are mutually orthogonal, $H_j H_k = 0$ for $j < k$. The second scenario considers the other extreme situation; the candidate models are nested, $H_j H_k = H_j$ for $j < k$. The more general situations where $H_j H_k$ is not a projection matrix are hard to derive analytic results.

Assume that $y_i = \mu_i + \varepsilon_i$, $E[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and put in the vector form $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$. Then the KL distance $L(\mathbf{w})$ is equivalent to the squared error loss, $\|\boldsymbol{\mu} - H(\mathbf{w})\mathbf{y}\|^2$ up to a proportionality constant, $2\sigma^2$, which will be dropped from the risk term:

$$R(\mathbf{w}) = \boldsymbol{\mu}'\bigl(I - H(\mathbf{w})\bigr)\bigl(I - H(\mathbf{w})\bigr)\boldsymbol{\mu} + \sigma^2 \mathbf{w}' H(\mathbf{w}) H(\mathbf{w})\mathbf{w}.$$

Consider the first scenario. Due to orthogonal projection, the bias term and the variance term of $R(\mathbf{w})$ can be computed from each model separately for each model and then combined together, yielding

$$R(\mathbf{w}) = \left\| \left(I - \sum_{k=1}^{M} H_k \right) \boldsymbol{\mu} \right\|^2 + \sum_{k=1}^{M} \bigl[(1 - w_k)^2 \|H_k \boldsymbol{\mu}\|^2 + w_k^2 p_k \sigma^2 \bigr].$$

Because of the weight relaxation, the minimization can be taken term by term, leading to the optimal weight assignment

$$w_k = \|H_k\boldsymbol{\mu}\|^2 / \bigl(\|H_k\boldsymbol{\mu}\|^2 + p_k\sigma^2\bigr) = \mathrm{SNR}_k / (1 + \mathrm{SNR}_k),$$

where we define the signal over noise ratio (SNR) for the kth model as $\mathrm{SNR}_k = \|H_k\boldsymbol{\mu}\|^2/(p_k\sigma^2)$. This indicates that the optimal weight on each model is between 0 and 1 and the sum can take any value between 0 and M.

To investigate the second scenario, we define the projection matrix $Q_k = H_k - H_{k-1}$, $\mathrm{Rank}(Q_k) = q_k = p_k - p_{k-1}$ for $k = 2, \dots, M$ and $Q_1 = H_1$, $\mathrm{Rank}(Q_1) = q_1 = p_1$. Then we have $Q_j Q_k = 0$ for $j \neq k$. Define $\tilde{w}_k = \sum_{i=k}^{M} w_i$ and $q_k = p_k - p_{k-1}$. We can rewrite $R(\mathbf{w})$ in terms of Q_k and $\tilde{w}_k$:

$$R(\mathbf{w}) = \|(I - H_M)\boldsymbol{\mu}\|^2 + \sum_{k=1}^{M} [(1 - \tilde{w}_k)^2 \|Q_k \boldsymbol{\mu}\|^2 + \tilde{w}_k^2 q_k \sigma^2].$$

Therefore, the optimal weights can be obtained from

$$\tilde{w}_k = \|Q_k \boldsymbol{\mu}\|^2 / (\|Q_k \boldsymbol{\mu}\|^2 + q_k \sigma^2),$$

leading to $w_k = \tilde{w}_k - \tilde{w}_{k+1}$. Note that to satisfy the nonnegativity constraint on w_k, the signal over noise ratio $\|Q_k \boldsymbol{\mu}\|^2 / q_k \sigma^2$ must be nonincreasing in k. The sum of optimal weights is equal to $\tilde{w}_1$, which is between 0 and 1.

5. Numerical results.

Because the performance of model averaging may depend on the class of models for averaging, two versions of implementing Step 1 in Section 3 are used for obtaining the numerical results reported in this section. Following Remark 2, let T be the number of predictors with p-value smaller than the 5% level. We further set the number of predictors p_k in each model to be the same (i.e., $p^* = p_1 = p_2 = \cdots = p_M$).

The first version MCV1 used a pre-specified p^* and M, while the second version MCV2 attempted to optimize the choice of p^* and M subject to the constraint that the total number of predictors from all M models $Mp^* \leq T$.

5.1. *Simulation study.*

In this simulation study, we use a high-dimensional logistic regression model. Following the settings of Bühlmann, Kalisch and Maathuis (2010) and Ando and Li (2014), we generate p predictors from multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$ with $s_{ij} = \rho^{|i-j|}$. The true model is

$$\log\left(\frac{P(y_i = 1|\mathbf{x}_i)}{1 - P(y_i = 1|\mathbf{x}_i)}\right) = \sum_{j=1}^{p} \gamma_j x_{ji}, \qquad i = 1, \dots, n, \tag{5.1}$$

where $P(y_i = 1|\mathbf{x}_i)$ is the true conditional probability, p is the number of predictors and γ_j are the regression coefficients. Let s be the number of γ_j with $\gamma_j \neq 0$. The predictors with nonzero γ_j are called the true predictors. We generate these nonzero γ_j from standard normal distribution $N(0, 1)$.

5.1.1. *Data generating process.*

Five settings with varying n, p, s and ρ are considered. In the first four settings, we implement the proposed method under the correctly specified models. These four settings are described as follows:

(a) Set the sample size $n = 100$, the number of predictors $p = 1000$ and $\rho = 0$. We set the number of true predictors $s = 40$, and let the nonzero γ_j be evenly spaced, $j = 10(h - 1) + 1$, $h = 1, \dots, 40$.

(b) Under the setting (a), the number of predictors is increased to $p = 2000$.

(c) Under the setting (b), the correlation parameter is changed to $\rho = 0.6$.

(d) Under the setting (a), the value of ρ is increased to $\rho = 0.9$, the number of predictors is increased to $p = 4000$, the number of observations is increased to $n = 200$.

For the fifth setting, we investigate the performance under model misspecification. Similar to Zhang, Li and Tsai (2010) and Flynn et al. (2013), the setting (e) studies the situation where some true predictors are excluded from the dataset.

(e) Set the number of true predictors $s = 70$ and let the location of true predictors γ_j be evenly spaced, $j = 10(h - 1) + 1$, $h = 1, \dots, 70$. Set $\rho = 0.9$, $n = 100$, generate 2020 predictors x_{ji} and the response variable y_i from (5.1). Then delete 20 true predictors x_{ji} for $j = 10(h - 1) + 1$, $h = 1, \dots, 20$ from the dataset. Use $2000(= p)$ remaining predictors for prediction.

5.1.2. *Implementation.* For MCV1, we set $p^* = 5$ and $M = 10$ to yield a class of 10 models, each with 5 predictors. This class of models is also used in implementing the traditional model averaging that requires the usual constraint of weights summing to 1, AIC model averaging (MAIC) and BIC model averaging (MBIC).

For MCV2, we choose p^* and M by minimizing the objective function $CV(\mathbf{w})$ subject to $p^* M \leq T$ for $p^* = 5, 10$. Note that the number of significant predictors T varies in each simulation. The average of T over 100 simulation run is; 55 [for setting (a)], 100 [for setting (b)], 106 [for setting (c)], 290 [for setting (d)] and 120 [for setting (e)].

To implement the MCP [Breheny and Huang (2011), Zhang (2010)] and SCAD [Fan and Li (2001)] algorithms, we used the R package ncvreg. To select an optimal size of penalty, we performed k-fold cross-validation for these penalized regression models over a grid of values for the regularization parameter. For this purpose, we implemented cv.ncvreg with default settings. The default value of $k = 10$ is used.

We also considered the original lasso [Tibshirani (1996)] and group lasso [Yuan and Lin (2006)] methods. We implemented the lasso logistic regression using the glmnet package in R. To select an optimal size of penalty, we performed cross-validation. For this purpose, we implemented cv.glmnet with the given default settings. For implementing group lasso, we partitioned the predictors into $M + 1$ groups. The first M groups are the same as those obtained with MAIC and MBIC. The last group consists of all the remaining predictors. We used the R package grplasso and used the $k(= 5)$-fold cross-validation procedure. A set of candidate

TABLE 1

The computational time required for each method. After 100 *simulation runs, the averaged time (in seconds) and corresponding standard deviations (SD) are given. MAIC, model averaging with the Akaike information criterion; MCV*1, *proposed model averaging with* $(M, p^*) = (10, 5)$; *MCV*2, *proposed model averaging and the number of models* M *and the number of predictors in each model are optimized; SCAD, penalized regression by SCAD approach; Lasso, original lasso procedure; MCP, panelized regression by MCP approach; G-Lasso, group lasso procedure*

Design	**MAIC**	**MCV1**	**MCV2**	**SCAD**	**Lasso**	**MCP**	**G-Lasso**
(a)	5.609	10.965	12.525	12.980	3.536	9.181	14.726
SD	0.033	0.058	0.174	0.376	0.057	0.282	0.087
(b)	10.652	15.827	16.926	15.103	4.594	10.519	27.496
SD	0.103	0.150	0.246	0.457	0.074	0.325	0.279
(c)	10.770	15.954	17.954	15.686	4.826	11.046	27.995
SD	0.108	0.149	0.281	0.454	0.097	0.336	0.264
(d)	25.308	37.760	114.447	34.982	20.253	16.970	67.861
SD	0.227	0.331	3.399	0.636	0.317	0.390	0.595
(e)	11.332	16.790	34.110	17.045	7.419	11.354	30.417
SD	0.060	0.073	0.607	0.378	0.100	0.321	0.137

values of the regularization parameter was prepared by using the function lambdamax in the R package grplasso.

5.1.3. *Results*. Table 1 compares the computational time required for each method, with the data sets generated. Here, we repeat the simulation 100 times and record the computational time for each. The table provides the mean (in seconds) and corresponding standard error of the calculated mean for each method. Given that the required computational times for MAIC and MBIC are identical, we just report the computational time required for MAIC. We can see that the computational time required for our method is not demanding.

We calculated the mean squared error (MSE) [average squared difference between the true $\sum_{j=1}^{p} \gamma_j x_{ji}$ in (5.1) and the estimates from each of the methods] as the performance measure for each method. Table 2 shows the averaged MSEs (and their standard errors) after 100 simulation runs. Evidently, the proposed model-averaging approach yields better performance than others.

It is further noted that the traditional model averaging procedures (AIC model averaging and BIC model averaging) can be implemented under the different settings of M and p^*. By varying (M, p^*), our method still performs favorably. Reports on these additional results are given in the supplementary document.

5.2. *Real data analysis*. Business analytics is often helpful to increase the efficiency of marketing campaigns. Moro, Laureano and Cortez (2011) studied a Portuguese marketing campaign related to bank deposit subscription with the goal of identifying a model that could explain the success of getting a client to subscribe.

TABLE 2

The performance measure MSE and its standard deviation (SD) under various simulation designs. MAIC, model averaging with the Akaike information criterion; MBIC, model averaging with the Bayesian information criterion; MCV1, proposed model averaging with $(M, p^*) = (10, 5)$; MCV2, proposed model averaging and the number of models M and the number of predictors in each model are optimized; SCAD, penalized regression by SCAD approach; Lasso, original lasso procedure; MCP, panelized regression by MCP approach; G-Lasso, group lasso procedure

Design	MAIC	MBIC	MCV1	MCV2	SCAD	Lasso	MCP	G-Lasso
(a)	29.940	29.940	20.759	20.655	35.201	35.560	37.046	34.230
SD	0.878	0.878	0.402	0.551	1.084	1.099	1.123	0.999
(b)	29.337	29.337	25.175	20.964	35.060	36.430	36.919	32.859
SD	0.832	0.832	0.575	0.573	1.077	1.095	1.008	0.896
(c)	27.814	27.814	22.815	20.280	32.557	32.196	34.280	31.529
SD	0.743	0.743	0.553	0.524	0.882	0.980	0.921	0.924
(d)	29.619	29.619	22.755	15.917	26.315	24.888	27.110	31.183
SD	0.919	0.919	0.747	0.491	0.922	0.848	0.899	0.981
(e)	54.637	54.637	43.039	35.933	58.799	57.626	61.073	58.738
SD	1.149	1.149	1.033	0.859	1.271	1.198	1.302	1.248

In this section, we apply the proposed model averaging method to the dataset from Moro, Laureano and Cortez (2011).

We analyzed 764 observations without missing covariates. The set of 15 variables under our study is summarized in the Table 3. The first six variables are numerical. We used these variables directly. The remaining variables ($x_7 \sim x_{15}$) are categorical (or binary). We transformed this information into indicator variables. After this operation, the total number of predictors is $p = 37$.

To evaluate the prediction performance, we randomly selected $n = 100$ observations for model fitting and used the rest of the data as the test set. Then we calculated the expected log-likelihood function

$$EL = \sum_{i=1}^{n_{\text{test}}} \bigl[z_i \log \hat{\pi}(\mathbf{x}_i) + (1 - z_i) \log\bigl(1 - \hat{\pi}(\mathbf{x}_i)\bigr)\bigr],$$

where $n_{\text{test}} = 664$ is the number of observations in the test set, z_i is the output of the ith observation in the test set, and $\hat{\pi}(\mathbf{x}_i) = E[z_i = 1|\mathbf{x}_i]$ is the predicted probability from the model. We repeated this process 1000 times to obtain the distribution of the EL values.

For MAIC, MBIC and MCV1, we set $M = 6$ and $p^* = p_1 = \cdots = p_6 = 5$. For the MCV2 procedure, we set the number of predictors $p^* = p_k = 4, 8, 12$ and let M vary freely. We selected the optimal (p^*, M) based on cross-validation. For implementing the lasso, grouped lasso, SCAD and MCP procedures, we employed the same approach described in Section 5.1. For grouped lasso, we attempted to use the same partition method to generate groups as described earlier in the simula-

TABLE 3

Our study comprises a set of 15 *variables*, *of which the first six variables* ($x_1 \sim x_6$) *are numerical variables and the remaining nine variables* ($x_7 \sim x_{15}$) *are categorical or binary variables*

$\boldsymbol{x}$	**Name (Descriptions)**
x_1	Age
x_2	Average yearly balance (in euros)
x_3	Last contact duration (in seconds)
x_4	Number of contacts performed during this campaign and for this client
x_5	Number of days that passed by after the client was last contacted from a previous campaign
x_6	Number of contacts performed before this campaign and for this client
x_7	Type of job (categorical “administration”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar”, “services”, “self-employed”, “retired”, “technician”)
x_8	Marital status (“married”, “divorced”, “single”)
x_9	Education (“secondary”, “primary”, “tertiary”)
x_{10}	Has credit in default? (“yes”, “no”)
x_{11}	Has housing loan? (“yes”, “no”)
x_{12}	Has personal loan? (“yes”, “no”)
x_{13}	Contact communication type (“telephone”, “cellular”)
x_{14}	Last contact month of year (“January”, . . . , “December”)
x_{15}	Outcome of the previous marketing campaign (“other”, ”failure“, ”success“)

tion study. Figure 1 plots the expected log-likelihood function for the testing data. Compared with other procedures, MCV2 has higher median and smaller deviation.

6. Conclusion.

In this paper, we investigated how to extend model averaging from linear to nonlinear regression under the high-dimensional settings. Following Ando and Li (2014), we allowed the weights to vary freely between 0 and 1 without the usual constraint of summing up to 1. We derived proper conditions for the proposed leave-one-out cross validation to behave optimally in the sense of achieving the best infeasible risk bound ζ_n (an oracle type of bound) asymptotically. We considered generalized linear models in this paper and used the Kullback–Leibler distance as the risk measure in replacement of the squared error used in linear regression. We also resolved the important issue concerning the existence of pseudo true parameter for each candidate model.

The critical condition (A.3) puts a ceiling on M, the number of candidate models to average. While M is allowed to grow as the sample size n increases, the rate is modulated by ζ_n. As Ando and Li (2014) argued, due to the complex nature of high-dimensional regression, we would expect ζ_n to increase at a rate faster than root n [for regression with dimension > 4, derived from the universal optimal rate of Stone (1982)], implying that ζ_n^{2K}/n^K must tend to the infinity. Putting together,

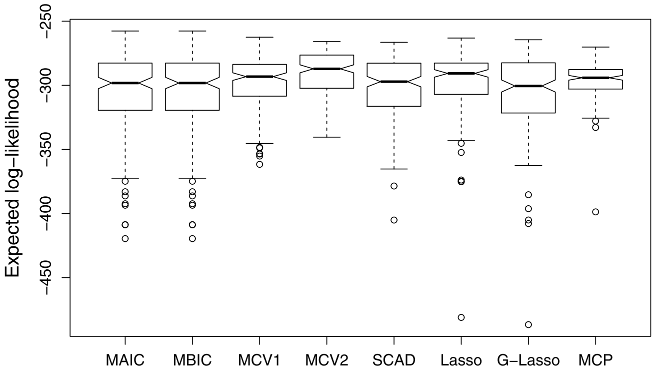

FIG. 1. *Boxplots of the expected log-likelihood function based on the test set*: $EL = \sum_{i=1}^{n_{\text{test}}}[z_i \log \hat{\pi}(\mathbf{x}_i) + (1 - z_i)\log(1 - \hat{\pi}(\mathbf{x}_i))]$, *where* $n_{\text{test}} = 664$ *is the number of observations in the test set*, z_i *is the output of the* i*th observation in the test set*, *and* $\hat{\pi}(\mathbf{x}_i) = E[z_i = 1|\mathbf{x}_i]$ *is the predicted probability from the model. MAIC*, *model averaging with the Akaike information criterion*; *MBIC*, *model averaging with the Bayesian information criterion*; *MCV*1, *model averaging without the restriction* $\sum_{k=1}^{M} w_k = 1$ *with* $(M, p^*) = (6, 5)$; *MCV*2, *model averaging without the restriction* $\sum_{k=1}^{M} w_k = 1$ *and the number of models* M *and the number of predictors in each model are optimized*; *SCAD*, *penalized regression by the SCAD approach*; *Lasso*, *original lasso procedure*; *G-Lasso*, *group lasso procedure*; *MCP*, *panelized regression by MCP approach.*

by ignoring the factor of $1/K$, the number of candidate models M is allowed to grow at the order no faster than $n^{1/4-\delta/2}$, where $1/2 > \delta > 0$.

One referee made a keen observation on the possibility of further relaxing the weights to allow for negative weights. By inspecting the proof of Theorem 1, it is confirmed that Theorem 1 still holds if conditions (4.3) and (4.4) are validate under the new weight space $Q_n = [-1, 1]^M$. However, for the simulation settings employed in this paper, the optimal weights were reached at positive values. Because the prediction by each individual model is positively correlated with the output y, negative weights appear less intuitive. This issue deserves further investigation.

Another issue is whether asymptotically it is necessary to relax the traditional stringent weight constraint. While it is difficult to obtain general analytic results, our preliminary study in Section 4.4 considered two scenarios of preparing candidate models, the mutually-orthogonal class and the nested-model class. Our results indicate that not only the positive weight relaxation but, sometimes, the allowance for negative weights can also lead to a smaller risk.

APPENDIX

PROOF OF LEMMA 1. Consider the case $\Theta = (-\infty, \infty)$ first. Because $b'(\theta)$ is increasing in θ, we have $b'(\infty) = \lim_{\theta\to\infty} b'(\theta)$ and $b'(-\infty) = \lim_{\theta\to-\infty} b'(\theta)$. $b'(\infty)$ can be finite or $+\infty$, and similarly, $b'(-\infty)$ can be finite or $-\infty$. Because of the convexity of $b(\cdot)$, it suffices to show that

$$\frac{d}{dc}G(\boldsymbol{\beta}_0 + c\mathbf{e}) = \sum_{i=1}^{n} \mu_i \mathbf{x}_i'\mathbf{e} - \sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} \tag{A.1}$$

as an increasing function of c is zero-crossing. This is obvious because

$$\begin{aligned}
&\lim_{c\to\infty} \sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} \\
&\quad = \lim_{c\to\infty} \sum_{\mathbf{x}_i'\mathbf{e}>0} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} + \lim_{c\to\infty} \sum_{\mathbf{x}_i'\mathbf{e}<0} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} \\
&\quad = \sum_{\mathbf{x}_i'\mathbf{e}>0} b'(\infty)\mathbf{x}_i'\mathbf{e} + \sum_{\mathbf{x}_i'\mathbf{e}<0} b'(-\infty)\mathbf{x}_i'\mathbf{e} \\
&\quad > \sum_{i=1}^{n} \mu_i \mathbf{x}_i'\mathbf{e}
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\lim_{c\to-\infty} \sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} \\
&\quad = \sum_{\mathbf{x}_i'\mathbf{e}>0} b'(-\infty)\mathbf{x}_i'\mathbf{e} + \sum_{\mathbf{x}_i'\mathbf{e}<0} b'(\infty)\mathbf{x}_i'\mathbf{e} \\
&\quad < \sum_{i=1}^{n} \mu_i \mathbf{x}_i'\mathbf{e}.
\end{aligned}$$

Therefore, Lemma 1 is proved for $\Theta = (-\infty, \infty)$.

Now, consider the case of the finite interval $\Theta = (\theta_1, \theta_2)$. In order to have $\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e} \in \Theta$, we must have

$$(\theta_1 - \mathbf{x}_i'\boldsymbol{\beta}_0)/(\mathbf{x}_i'\mathbf{e}) < c < (\theta_2 - \mathbf{x}_i'\boldsymbol{\beta}_0)/(\mathbf{x}_i'\mathbf{e}) \qquad \text{for } \mathbf{x}_i'\mathbf{e} > 0$$

and

$$(\theta_2 - \mathbf{x}_i'\boldsymbol{\beta}_0)/(\mathbf{x}_i'\mathbf{e}) < c < (\theta_1 - \mathbf{x}_i'\boldsymbol{\beta}_0)/(\mathbf{x}_i'\mathbf{e}) \qquad \text{for } \mathbf{x}_i'\mathbf{e} < 0.$$

Note that $\mathbf{x}_i'\boldsymbol{\beta}_0 \in \Theta$ implies $\theta_2 - \mathbf{x}_i'\boldsymbol{\beta}_0 > 0$ and $\theta_1 - \mathbf{x}_i'\boldsymbol{\beta}_0 < 0$. The allowable set of c is the intersection of these intervals. Let (c_1, c_2) be the intersection.

Another key observation is that $\Theta = (\theta_1, \theta_2)$ implies that

$$\lim_{\theta\to\theta_2^-} b(\theta) = \infty = \lim_{\theta\to\theta_1^+} b(\theta),$$

which in turn implies that

$$\lim_{\theta\to\theta_2^-} b'(\theta) = \infty \quad \text{and} \quad \lim_{\theta\to\theta_1^+} b'(\theta) = -\infty.$$

Using these equations, we can show that

$$\lim_{c\to c_2^-}\sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} = +\infty$$

and

$$\lim_{c\to c_1^+}\sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} = -\infty.$$

Therefore, the expression (A.1), as an increasing function of c, is zero-crossing.

The case that $(-\infty, \theta_1)$, for finite θ_1, can be proved by observing that

$$\lim_{\theta\to\theta_1^-} b(\theta) = \infty \quad \text{and} \quad \lim_{\theta\to\theta_1^-} b'(\theta) = \infty.$$

The allowable set of c is the intersection of

$$c < (\theta_1 - \mathbf{x}_i'\boldsymbol{\beta}_0)/(\mathbf{x}_i'\mathbf{e}) \qquad \text{for } \mathbf{x}_i'\mathbf{e} > 0$$

and

$$c > (\theta_1 - \mathbf{x}_i'\boldsymbol{\beta}_0)/(\mathbf{x}_i'\mathbf{e}) \qquad \text{for } \mathbf{x}_i'\mathbf{e} < 0.$$

Let (c_1, c_2) be the intersection of these intervals. Suppose $\mathbf{x}_i'\mathbf{e} > 0$ for some i. Then we can show that $\lim_{c\to c_2^-}\sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} = +\infty$.

Suppose $\mathbf{x}_i'\mathbf{e} < 0$ for all i; then $c_2 = +\infty$.

$$\lim_{c\to c_2^-}\sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} = \sum_{i=1}^{n} b'(-\infty)\mathbf{x}_i'\mathbf{e} > \sum_{i=1}^{n}\mu_i\mathbf{x}_i'\mathbf{e}.$$

Therefore, we can show that

$$\lim_{c\to c_2^-}\sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} > \sum_{i=1}^{n}\mu_i\mathbf{x}_i'\mathbf{e}.$$

Suppose $\mathbf{x}_i'\mathbf{e} < 0$ for some i; then, c_1 is finite.

$$\lim_{c\to c_1^+}\sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} = -\infty.$$

On the other hand, if $\mathbf{x}_i'\mathbf{e} > 0$ for all i, then $c_1 = -\infty$.

$$\lim_{c\to -\infty}\sum_{i=1}^{n} b'(\mathbf{x}_i'\boldsymbol{\beta}_0 + c\mathbf{x}_i'\mathbf{e})\mathbf{x}_i'\mathbf{e} = \sum_{i=1}^{n} b'(-\infty)\mathbf{x}_i'\mathbf{e} < \sum_{i=1}^{n}\mu_i\mathbf{x}_i'\mathbf{e}.$$

Therefore, we have proved the case $\Theta = (-\infty, \theta_1)$. The remaining case $\Theta = (\theta_1, \infty)$ can be proved in a similar manner. □

PROOF OF LEMMA 2. Let $\mathbf{e}_0$ be any vector in R^p with $\|\mathbf{e}_0\| = 1$. Using the epsilon-delta argument, for $\varepsilon > 0$, we want to find a $\delta > 0$ such that for any $\mathbf{e}$ with $\|\mathbf{e}\| = 1$ and $\|\mathbf{e} - \mathbf{e}_0\| = \delta$, the value of $c(\mathbf{e})$ satisfies

$$\left|c(\mathbf{e}) - c(\mathbf{e}_0)\right| < \varepsilon. \tag{A.2}$$

Denote $M_1 = \boldsymbol{\beta}_0 + c(\mathbf{e}_0)\mathbf{e}_0\text{m}$ and choose a small ε_1, $0 < \varepsilon_1 < \varepsilon$ so that both $M_2 = \boldsymbol{\beta}_0 + (c(\mathbf{e}_0) + \varepsilon_1)\mathbf{e}_0$ and $M_3 = \boldsymbol{\beta}_0 + (c(\mathbf{e}_0) - \varepsilon_1)\mathbf{e}_0$ fall into the allowable set B.

The uniqueness of $c(\mathbf{e}_0)$ implies that

$$G(M_1) - G(M_2) > 0 \quad \text{and} \quad G(M_1) - G(M_3) > 0.$$

Using the continuity of $G(\cdot)$, there exists a small $\delta > 0$ such that for any $\mathbf{e}$ with $\|\mathbf{e}\| = 1$ and $\|\mathbf{e} - \mathbf{e}_0\| = \delta$, the following three inequalities hold:

$$\begin{aligned}
\left|G(\boldsymbol{\beta}_0 + c(\mathbf{e}_0)\mathbf{e}) - G(M_1)\right| &< \min\bigl\{\bigl(G(M_1) - G(M_2)\bigr)/3, \\
&\qquad \bigl(G(M_1) - G(M_3)\bigr)/3\bigr\} \\
\left|G\bigl(\boldsymbol{\beta}_0 + \bigl(c(\mathbf{e}_0) + \varepsilon_1\bigr)\mathbf{e}\bigr) - G(M_2)\right| &< \bigl(G(M_1) - G(M_2)\bigr)/3, \\
\left|G\bigl(\boldsymbol{\beta}_0 + \bigl(c(\mathbf{e}_0) - \varepsilon_1\bigr)\mathbf{e}\bigr) - G(M_3)\right| &< \bigl(G(M_1) - G(M_3)\bigr)/3.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&G(\boldsymbol{\beta}_0 + c(\mathbf{e}_0)\mathbf{e}) - G\bigl(\boldsymbol{\beta}_0 + \bigl(c(\mathbf{e}_0) + \varepsilon_1\bigr)\mathbf{e}\bigr) \\
&\quad = G(M_1) - G(M_2) + \bigl\{G(\boldsymbol{\beta}_0 + c(\mathbf{e}_0)\mathbf{e}) - G(M_1)\bigr\} \\
&\qquad + \bigl\{G(M_2) - G(\boldsymbol{\beta}_0 + \bigl(c(\mathbf{e}_0) + \varepsilon_1\bigr)\mathbf{e}\bigr\} \\
&\quad \ge G(M_1) - G(M_2) - \bigl(G(M_1) - G(M_2)\bigr)/3 - \bigl(G(M_1) - G(M_2)\bigr)/3 \\
&\quad = \bigl(G(M_1) - G(M_2)\bigr)/3 > 0.
\end{aligned}$$

Similarly, we can derive $G(\boldsymbol{\beta}_0 + c(\mathbf{e}_0)\mathbf{e}) - G(\boldsymbol{\beta}_0 + (c(\mathbf{e}_0) - \varepsilon_1)\mathbf{e}) > 0$. Now, by the concavity of $G(\cdot)$, it is clear that $c(\mathbf{e})$ must fall between $c(\mathbf{e}_0) - \varepsilon$ and $c(\mathbf{e}_0) + \varepsilon$, providing the inequality (A.2). The proof of Lemma 2 is now complete. □

Acknowledgments. We thank the Co-Editor, an Associate Editor, two anonymous referees for helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/17-AOS1538SUPP; .pdf). Due to space constraints, the proof of the claims (4.8) and (4.9), the proof of Lemma 3, and further simulation studies are relegated to the supplementary document. Supplementary document also contains Theorem 3 and Lemma 4.

REFERENCES

AKAIKE, H. (1978). On the likelihood of a time series model. *J. R. Stat. Soc., Ser. D Stat.* **27** 217–235.

AKAIKE, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* **66** 237–242. MR0548189

ANDO, T. (2009). Bayesian portfolio selection using multifactor model. *Int. J. Forecast.* **25** 550–566.

ANDO, T. and LI, K.-C. (2014). A model-averaging approach for high-dimensional regression. *J. Amer. Statist. Assoc.* **109** 254–265. MR3180561

ANDO, T. and LI, K. (2017). Supplement to "A weight-relaxed model averaging approach for high-dimensional generalized linear models." DOI:10.1214/17-AOS1538SUPP.

ANDO, T. and TSAY, R. (2010). Predictive likelihood for Bayesian model selection and averaging. *Int. J. Forecast.* **26** 744–763.

BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396

BÜHLMANN, P., KALISCH, M. and MAATHUIS, M. K. (2010). Variable selection in high-dimensional linear models: Partially faithful distributions and the PC-simple algorithm. *Biometrika* **97** 261–278.

CHARKHI, A., CLAESKENS, G. and HANSEN, B. E. (2016). Minimum mean squared error model averaging in likelihood models. *Statist. Sinica* **26** 809–840.

CHUNG, T. S., RUST, R. T. and WEDEL, M. (2009). My mobile music: An adaptive personalization system for digital audio players. *Marketing Sci.* **28** 52–68.

CLAESKENS, G. and HJORT, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98** 900–945. MR2041482

EKLUND, J. and KARLSSON, S. (2007). Forecast combination and model averaging using predictive measures. *Econometric Rev.* **26** 329–363. MR2364365

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

FLYNN, C. J., HURVICH, C. M. and SIMONOFF, J. S. (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *J. Amer. Statist. Assoc.* **108** 1031–1043.

HANSEN, B. E. (2007). Least squares model averaging. *Econometrica* **75** 1175–1189. MR2333497

HANSEN, B. E. and RACINE, J. S. (2012). Jackknife model averaging. *J. Econometrics* **167** 38–46. MR2885437

HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98** 879–899. MR2041481

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176

KASS, R. and RAFTERY, A. (1995). Bayes factors and model uncertainty. *J. Amer. Statist. Assoc.* **90** 773–795.

LEE, Y. S. (2014). Management of a periodic-review inventory system using Bayesian model averaging when new marketing efforts are made. *Int. J. Production Econ.* **158** 278–289.

LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1011–1112.

LI, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** 958–975.

LV, J. and LIU, J. S. (2014). Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 141–167.

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MIN, K.-C. and ZELLNER, A. (1992). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *J. Econometrics* **56** 89–118.

MONTGOMERY, J. M. and NYHAN, B. (2010). Bayesian model averaging: Theoretical developments and practical applications. *Polit. Anal.* **18** 245–270.

MORO, S., LAUREANO, R. and CORTEZ, P. (2011). Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In *Proceedings of the European Simulation and Modelling Conference* 117–121.

OUYSSE, R. and KOHN, R. (2010). Bayesian variable selection and model averaging in the arbitrage pricing theory model. *Comput. Statist. Data Anal.* **54** 3249–3268.

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92** 179–191.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288.

WAN, A. T. K., ZHANG, X. and ZOU, G. (2010). Least squares model averaging by Mallows criterion. *J. Econometrics* **156** 277–283. MR2609932

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25.

YEUNG, K. E., BUMGARNER, R. E. and RAFTERY, A. E. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* **21** 2394–2402.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, Y., LI, R. and TSAI, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. MR2656055

MELBOURNE BUSINESS SCHOOL
UNIVERSITY OF MELBOURNE
CARLTON, VICTORIA 3053
AUSTRALIA
E-MAIL: T.Ando@mbs.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, LOS ANGELES
LOS ANGELES, CALIFORNIA 90095
USA
AND
INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIPEI 11529
TAIWAN
E-MAIL: kcli@stat.ucla.edu
kcli@stat.sinica.edu.tw