




A Model-Averaging Approach for High-Dimensional Regression

Tomohiro Ando & Ker-Chau Li


To cite this article: Tomohiro Ando & Ker-Chau Li (2014) A Model-Averaging Approach for High-Dimensional Regression, Journal of the American Statistical Association, 109:505, 254-265, DOI: [10.1080/01621459.2013.838168](https://doi.org/10.1080/01621459.2013.838168)

To link to this article: <https://doi.org/10.1080/01621459.2013.838168>

 View supplementary material 

 Accepted author version posted online: 26 Sep 2013.
Published online: 19 Mar 2014.

 Submit your article to this journal 

 Article views: 2290

 View related articles 

 View Crossmark data 

 Citing articles: 47 View citing articles 

A Model-Averaging Approach for High-Dimensional Regression

Tomohiro ANDO and Ker-Chau LI

This article considers high-dimensional regression problems in which the number of predictors p exceeds the sample size n . We develop a model-averaging procedure for high-dimensional regression problems. Unlike most variable selection studies featuring the identification of true predictors, our focus here is on the prediction accuracy for the true conditional mean of y given the p predictors. Our method consists of two steps. The first step is to construct a class of regression models, each with a smaller number of regressors, to avoid the degeneracy of the information matrix. The second step is to find suitable model weights for averaging. To minimize the prediction error, we estimate the model weights using a delete-one cross-validation procedure. Departing from the literature of model averaging that requires the weights always sum to one, an important improvement we introduce is to remove this constraint. We derive some theoretical results to justify our procedure. A theorem is proved, showing that delete-one cross-validation achieves the lowest possible prediction loss asymptotically. This optimality result requires a condition that unravels an important feature of high-dimensional regression. The prediction error of any individual model in the class for averaging is required to be higher than the classic root n rate under the traditional parametric regression. This condition reflects the difficulty of high-dimensional regression and it depicts a situation especially meaningful for $p > n$. We also conduct a simulation study to illustrate the merits of the proposed approach over several existing methods, including lasso, group lasso, forward regression, Phase Coupled (PC)-simple algorithm, Akaike information criterion (AIC) model-averaging, Bayesian information criterion (BIC) model-averaging methods, and SCAD (smoothly clipped absolute deviation). This approach uses quadratic programming to overcome the computing time issue commonly encountered in the cross-validation literature. Supplementary materials for this article are available online.

KEY WORDS: Asymptotic optimality; High-dimensional regression models; Model weights.

1. INTRODUCTION

Model selection and model averaging are two approaches used to improve prediction accuracy in the classical setting of regression analysis where the sample size is at least one order of magnitude greater than the number of regressors. Typically, a class of candidate models with varying degrees of model complexity is specified first. Well-known model selection procedures, such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and forward and backward regression yield one optimal model to use for prediction. Instead of relying on only one best model, the model-averaging approach advocates the pooling of predictions by giving higher weights to the better models.

In recent years, many studies have encountered datasets with a very large number of potential predictors, among which only very few are considered to be truly relevant in prediction. To identify informative predictors under this signal sparseness assumption, many novel variable procedures have been developed in the context of high-dimensional regression. These include the lasso method (Tibshirani 1996; Zou, Hastie, and Tibshirani 2007) and its variants (Yuan and Lin 2006; Zou 2006), Bayesian lasso (Park and Casella 2008), least-angle regression (Efron et al. 2004), elastic net (Zou and Hastie 2005), the smoothly clipped absolute deviation penalty (SCAD; Fan and Li 2001), the minimax concave penalty method (MCP; Breheny and Huang

2011), the Dantzig selector (Candes and Tao 2007), marginal regression (Genovese, Jin, and Wasserman 2009), correlation learning (Fan and Lv 2008), and the partial faithfulness approach (Buhlmann, Kalisch, and Maathuis 2010) among many others. The literature on this field of research continues to expand rapidly.

By contrast, very little attention has been paid to the problem of how to conduct model averaging in high-dimensional linear models, especially when the number of predictors greatly exceeds the sample size. So far, the major focus of model averaging has been on the determination of the weights for individual models under the standard setting in which the number of observations is greater than the number of predictors. These studies include forecast model averaging (Newbold and Granger 1974), AIC model averaging (Akaike 1979), BIC model averaging (Palm and Zellner 1992; Madigan and Raftery 1994; Kass and Raftery 1995; Raftery, Madigan, and Hoeting 1997; Hoeting et al. 1999; Fernandez, Ley, and Steel 2001), the Mallows C_p model averaging (Hansen 2007; Wan, Zhang, and Zou 2010; see also Mallows 1973), Bayesian model averaging using predictive measures (Eklund and Karlsson 2007), jackknife model averaging (Hansen and Racine 2012), and predictive likelihood model averaging (Ando and Tsay 2010). Computationally, the delete-one cross-validation (jackknife) method of weight determination by Hansen and Racine reduces to just a quadratic optimization problem, which can be solved by available quadratic programming packages. This is a great advantage in dealing with high-dimensional regression as we shall demonstrate in this article.

Tomohiro Ando is Associate Professor, Graduate School of Business Administration, Keio University, Kanagawa, Japan (E-mail: andoh@kbs.keio.ac.jp). Ker-Chau Li is Professor, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan and Department of Statistics, UCLA, CA 90095 (E-mail: kcli@stat.sinica.edu.tw). The authors thank the co-editor, the associate editor, and two anonymous reviewers for constructive and helpful comments that improved the quality of the article considerably. We are also grateful to them for providing several references. The research of TA is partially supported by Inamori Foundation, Japan. The work of KCL is supported in part by NSF grant DMS-0707160 and by internal funding of Academia Sinica.

We present a model-averaging method for high-dimensional regression. Our article has several contributions. First, the algorithm developed is computationally feasible even when there are thousands of covariates. Second, the standard restriction whereby the sum of the model weights is equal to one is relaxed. To the best of our knowledge, ours is the first study that removes this restriction. We show that this relaxation is important for improving the prediction performance. Third, the method is theoretically examined. We prove that the minimization of the cross-validation criterion asymptotically minimizes the squared error between the true mean and the predicted value, an oracle property similar to Li (1986, 1987) or Shao (1997) in the model selection context. Fourth, our theoretical result unravels an important distinction between our problem and the traditional parametric regression, which has not received much attention in the high-dimensional regression literature. The prediction error of any individual model in the class for averaging is required to be higher than the classic root n rate under the traditional parametric regression. This condition properly reflects the difficulty encountered with the increased dimensionality, especially for $p > n$. Finally, we propose a practical method for resolving the issue of how to set the number of models to be averaged. We assess the method's performance by simulation. The results indicate that our method yields more accurate prediction than many existing methods, including lasso, group lasso, partial faithfulness approach, AIC model-averaging, BIC model-averaging methods, and SCAD.

The remainder of the article is organized as follows. Section 2 describes the problem setting. A model-averaging procedure is introduced for high-dimensional linear models in which the number of predictors may exceed the sample size. We provide some theoretical results in Section 3. Section 4 presents simulation evidence to illustrate the merits of the proposed method. The performance is compared with that of previously proposed model-averaging procedures, lasso, and its variants. Further discussion and concluding remarks are given in Section 5.

2. PROBLEM SETTING AND METHOD

Suppose we have n independent observations $\{(y_\alpha, \mathbf{x}_\alpha); \alpha = 1, 2, \dots, n\}$, where y_α is the response variable and \mathbf{x}_α is a vector of p -dimensional explanatory variables. The number of predictors is allowed to increase with the sample size n . The problem to be considered is the quantification of the relationship between the response variable and the explanatory variables from observed data. We consider the multiple linear regression model

$$y = \sum_{j=1}^p \beta_j x_j + \varepsilon, \quad (1)$$

where the independent random error ε has mean $E[\varepsilon] = 0$ and finite variance that is allowed to be heterogenous. Here, we have dropped the intercept term by assuming that the means of the response variable and the input variables are already subtracted out.

Following the convention in model selection, we assume that not all predictors have contributions in predicting the response. We denote the number of true predictors (i.e., those with nonzero regression coefficients β_j) by s . Both s and the set of true predictors $T = \{j : |\beta_j| > 0\}$ are unknown. The elements of predic-

tors may incorporate terms in a series basis expansion, including splines, wavelets, kernels and so on.

Model averaging involves an attempt to combine several competing scientific models. Different models may be proposed by different research teams with possibly different ideas, computing systems, and background knowledge. Previous studies on model averaging have focused on situations in which the total number of predictors p is much smaller than the sample size n . Mathematically, the flexibility of varying the model weights may expand the prediction space, which is beneficial from a prediction point of view.

The present study addresses the situation in which the accuracy in predicting the mean of y is the primary concern. Our approach is mainly motivated by the attempt to address the dimensionality issue encountered in regression problems with $p > n$. There are two steps involved.

2.1 Step 1: Prepare the Candidate Models

We denote a set of M candidate models M_1, \dots, M_M by

$$M_k : y = \sum_{j \in A_k} \beta_j x_j + \varepsilon, \quad (2)$$

where A_k is the index set of regressors to be included in model M_k , $k = 1, \dots, M$. Each model can also be written in matrix form $y = X_k \beta_k + \varepsilon$, where β_k is the p_k -dimensional parameter vector and ε is the n -dimensional noise vector.

The noise elements ε_i may have different variance σ_i^2 , $i = 1, \dots, n$, which are unknown to us. We estimate the regression coefficients by the usual least-square method:

$$\hat{\beta}_k = \operatorname{argmin} \|y - X_k \beta_k\|^2,$$

which leads to $\hat{\beta}_k = (X_k' X_k)^{-1} X_k' y$ and the least-square prediction $\hat{\mu}_k = X_k \hat{\beta}_k$ for $k = 1, \dots, M$.

Depending on the context, the candidate models can be provided in various ways. In the fields of economics, finance, and business, there is a wide body of literature on model averaging for applications such as stock returns (Avramov 2002), exchange rates (Wright 2008), output growth (Garratt et al. 2003), portfolio management (Pesaran, Schleicher, and Zaffaroni 2009), currency crisis analysis (Crespo-Cuaresma and Slacik 2009), analysis of preferential trade agreements (Eicher, Henn, and Papageorgiou 2012), tourism studies (Wan and Zhang 2009), and health studies (Morales et al. 2006). Most of these studies assume that candidate models are based on different competing theories for prediction of the outcome variable y .

Here, we propose a method for preparing candidate models without prior subject knowledge or expert theories. The idea stems from the common perception that higher marginal correlation between an input variable and the output variable suggests better prediction power. Our proposal groups the regressors with similar size of correlation together to form models for averaging.

We first calculate the marginal correlation between each predictor variable and the response variable. We partition the p marginal correlations into $M + 1$ groups by their absolute values. The first group has the highest values and the $M + 1$ group has values closest to zero. Let model M_k consist of the regressors with marginal correlations falling into the k th group. We drop the $M + 1$ group. Thus the number of models is M .

Without loss of generality, we assume that the design matrix X has been standardized (i.e., for each j , the j th column of X satisfies $n^{-1}||X_{(j)}||^2 = n^{-1}X'_{(j)}X_{(j)} = 1$). Then the marginal correlation between each predictor variable and the response variable is estimated by $\hat{y} = n^{-1}X'y$. Sorting the set of p regressors based on the marginal correlation magnitude, we obtain M design matrices X_1, \dots, X_M for (2). Here, X_1 , with marginal correlation of the greatest magnitude, is used for model M_1 . A similar argument applies to X_2, \dots, X_M . The remaining regressors not included in X_1, \dots, X_M are dropped at this stage. Under certain conditions, we show that all true regressors are included in (2).

2.2 Step 2: Optimize the Model Weights

After a list of candidate models is specified and their least-square predictors $\{\hat{\mu}_1, \dots, \hat{\mu}_M\}$ are obtained, we have to determine the weight of each model. We denote the hat matrix $X_k(X'_k X_k)^{-1}X_k$ by H_k .

Let the M -dimensional weight vector $\mathbf{w} = (w_1, \dots, w_M)'$ come from the unit hypercube of R^M :

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

The model-average predictor $\hat{\mu}$ can be written as

$$\begin{aligned}\hat{\mu} &= \sum_{k=1}^M w_k \hat{\mu}_k = \sum_{k=1}^M w_k X_k (X'_k X_k)^{-1} X'_k y \\ &= \sum_{k=1}^M w_k H_k y = H(\mathbf{w})y,\end{aligned}$$

where $H(\mathbf{w}) = \sum_{k=1}^M w_k H_k$ is the corresponding hat matrix. Usually, the M -dimensional weight vector is restricted such that $\sum_{k=1}^M w_k = 1$. In this article, we remove this restriction.

We estimate the weights using the delete-one cross-validation approach as used in Hansen and Racine (2012). Let $\tilde{\mu}_k^{(-\alpha)}$ be the predicted value of the α th observation from the k th model M_k , which is estimated from the observations except for $(y_\alpha, \mathbf{x}_\alpha)$. Let $\tilde{\mu}_k = (\tilde{\mu}_k^{(-1)}, \dots, \tilde{\mu}_k^{(-n)})'$ be an n -dimensional vector. As shown in Li (1986), we can write $\tilde{\mu}_k = \tilde{H}_k y$, where \tilde{H}_k is the smoothing matrix given by $\tilde{H}_k = D_k(H_k - I) + I$ and D_k is the $n \times n$ diagonal matrix with the α th diagonal element equal to $(1 - h_{k\alpha})^{-1}$, where $h_{k\alpha}$ is the α th diagonal element of H_k . Then the delete-one predictor is

$$\tilde{\mu} = \sum_{k=1}^M w_k \tilde{\mu}_k = \sum_{k=1}^M w_k \tilde{H}_k y = \tilde{H}(\mathbf{w})y,$$

where $\tilde{H}(\mathbf{w}) = \sum_{k=1}^M w_k \tilde{H}_k$. We use the sum of squared residuals of the delete-one predictor to form the cross-validation criterion $CV(\mathbf{w}) = (\mathbf{y} - \tilde{\mu})'(\mathbf{y} - \tilde{\mu}) = (\mathbf{y} - \tilde{H}(\mathbf{w})\mathbf{y})'(\mathbf{y} - \tilde{H}(\mathbf{w})\mathbf{y})$. We select the weight vector \mathbf{w} that minimizes the delete-one cross-validation criterion over the set Q_n :

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in Q_n} CV(\mathbf{w}). \quad (3)$$

Unlike other cross-validation problems that are often time-consuming to compute, it is much easier for our problem. The global optimization can be performed by constrained optimization programming. For example, `optim` package in R language and other open software are applicable. For optimization under the restriction $\sum_{k=1}^M w_k = 1$, we may use `quadprog` package

in R language, `quadprog` command in MATLAB, `qprog` command in GAUSS, and other open software.

Why should the restriction $\sum_{k=1}^M w_k = 1$ be removed? There are at least three reasons. First, recall that the traditional model averaging is motivated by applications where all models are equally competitive. Fixing the total weight allows the data to determine the relative contribution of final prediction by each model objectively. However, the situation is quite different in our case. The models proposed in Step 1 do not seem to be equally competitive. The first few models are likely to be more informative than the last few models because of the ordering of regressors by marginal correlations with y . If we were to fix the total weight, then to shift the weights away from the first few models would be nonbeneficial for prediction.

Second, consider the extreme case that the regressors are uncorrelated with each other and the noise variance is ignorable. The predictors from each model become uncorrelated with each other as well and the optimal combined predictor is the sum of all model predictors, implying that the optimal weight assignment should be $(1, 1, \dots, 1)$. Thus, the total weight should be equal to M , not 1.

Third, we are able to prove the main theorem of this article (Theorem 1 in Section 3) without the total weight constraint. Thus, in general, relaxing the total weight constraint is likely to lower the prediction error.

Remark 2.1. For a set of grouped predictors, the group lasso (Yuan and Lin 2006) could be used. Meier, van de Geer, and Bühlmann (2008) pointed out that a consistent result can be obtained for the group lasso for Gaussian regression. But a critical issue is how to determine an optimal value of the regularization parameter. Yuan and Lin suggested a modified version of the C_p criterion. However, exact computation of the degrees of freedom is difficult. Although Yuan and Lin proposed the use of bootstrapping, it is a computationally intensive task. Moreover, like all C_p methods, the value of noise variance must be specified, which may introduce additional uncertainty harder to assess theoretically. By contrast, our proposed method at Step 2 does not suffer from these problems. Our method is computationally more efficient. We are also able to derive asymptotic optimality in Section 3, while no such oracle property has been derived for choosing penalty parameters in lasso-type of procedures.

Remark 2.2. As pointed out by a referee, the cross-validation criterion $CV(\mathbf{w})$ can be further expressed as

$$\begin{aligned}CV(\mathbf{w}) &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\tilde{H}(\mathbf{w})\mathbf{y} + \mathbf{y}'\tilde{H}(\mathbf{w})'\tilde{H}(\mathbf{w})\mathbf{y} \\ &= \mathbf{y}'\mathbf{y} - 2\sum_{k=1}^M w_k \mathbf{y}'\tilde{H}_k \mathbf{y} + \sum_{k=1}^M \sum_{m=1}^M w_k w_m \mathbf{y}'\tilde{H}_k' \tilde{H}_m \mathbf{y} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{w}\mathbf{a} + \mathbf{w}\mathbf{B}\mathbf{w},\end{aligned}$$

where \mathbf{a} is an M -dimensional vector with the k th element $a_k = \mathbf{y}'\tilde{H}_k \mathbf{y}$, and \mathbf{B} is an $M \times M$ dimensional matrix with (k, m) th element $B_{k,m} = \mathbf{y}'\tilde{H}_k' \tilde{H}_m \mathbf{y}$. Therefore, the cross-validation criterion $CV(\mathbf{w})$ is further expressed the quadratic form in \mathbf{w} ; and thus we can also use quadratic programming. Although Table 1 shows that the computational time required for the constrained optimization is also fast, the computational time required for minimizing $CV(\mathbf{w})$ can be further improved by using the quadratic programming.

Table 1. The computational time required for each method under the simulation design (a) with $n = 50$ and $n = 100$. After 100 simulation runs, the averaged time (in seconds) and corresponding standard deviations (SDs) are given. MAIC, model averaging with the Akaike information criterion under the restriction $\sum_{k=1}^M w_k = 1$; MCV1, cross-validation model averaging without the restriction $\sum_{k=1}^M w_k = 1$ and the number of models and the number of regressors in each model is fixed; MCV2, cross-validation model averaging without the restriction $\sum_{k=1}^M w_k = 1$ and the number of models M and the number of regressors in each model are optimized; SCAD, penalized regression by SCAD approach; Lasso, original lasso procedure; G-Lasso, group lasso procedure; MCP, panelized regression by MCP approach

n		MAIC	MCV1	MCV2	SCAD	Lasso	G-Lasso	MCP
50	Mean	0.037	0.076	0.492	10.292	0.674	221.605	8.459
	SD	0.016	0.033	0.049	1.718	0.056	11.320	0.611
100	Mean	0.053	0.107	0.701	18.409	1.485	242.799	17.663
	SD	0.008	0.016	0.093	1.603	0.107	18.054	1.770

3. THEORETICAL RESULTS

This section theoretically investigates some properties of the proposed model-averaging procedure. In Section 3.1, we prove that the cross-validation procedure is asymptotically optimal for the weight determination. The main result, Theorem 1, does not require that the regressors in different models must be nonoverlapping or be uncorrelated. It can be applied to classes of models prepared by many procedures. Section 3.2 concerns the ordering of regressors by marginal correlations with output variable.

3.1 Optimality of the Cross-Validation Criterion

Denote $X = \{x_1, \dots, x_n\}$ and $\mu = E(y|X)$. Define the sum of squared estimation errors of a model-averaging estimator $\hat{\mu} = \sum_{k=1}^M w_k H_k y$ in estimating μ by $L(w) = (\mu - \hat{\mu})'(\mu - \hat{\mu})$ and consider the risk of estimation $R(w) = E[L(w)|X] = E[(\mu - \hat{\mu})'(\mu - \hat{\mu})|X]$.

Intuitively cross-validation allows the data to determine the best weights to use. We shall make certain assumptions so that the difference between $\hat{\mu} = H(w)y$ and the delete-one cross-validation estimate $\tilde{H}(w)y$ is small asymptotically. We can show that the cross-validation criterion $CV(w)$ yields an unbiased estimate of the risk $R(w)$ up to a constant term independent of w asymptotically. This indicates that cross-validation almost behaves like a procedure that gives the lowest risk among all weight choices,

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

Theorem 1 presents an even stronger sense of optimality, which is similar to that shown in Li (1986, 1987). It shows that the smallest possible loss $\inf_{w \in Q_n} L(w)$, which is infeasible to achieve because the true μ is unknown, is indeed within the reach of cross-validation.

Theorem 1. As $n \rightarrow \infty$, assume that for some fixed integer $1 \leq K < \infty$,

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

$$\sup_k \frac{1}{p_k} \bar{\lambda}\{H_k\} \leq \Lambda n^{-1} \quad (5)$$

$$\sup_{1 \leq k \leq M} \frac{p_k}{n^{3/4}} \leq \Lambda' < \infty \quad (6)$$

$$M^{4K+2} \|\mu\|^{2K} / \zeta_n^{2K} \rightarrow 0, \quad (7)$$

$$0 < C_1 < \|\mu\|^2 / n < C_2 < \infty \text{ for some constant } C_1, C_2, \quad (8)$$

where B , Λ , and Λ' are finite constants, $\bar{\lambda}\{\cdot\}$ denotes the maximal diagonal element of a matrix, and p_k is the number of columns of X_k . Then, we have

$$\frac{L(\hat{w})}{\inf_{w \in Q_n} L(w)} \rightarrow 1, \quad (9)$$

where the convergence is in probability.

Remark on Theorem 1. We believe that Theorem 1 provides a theoretical foundation for the proposed model-averaging procedure described in Section 2. Condition (4) is a moment condition concerning the random errors and is satisfied by Gaussian noise. Condition (5) is the same as condition (5.2) of Li (1987). As pointed out in Li (1987), this condition excludes the case of extremely unbalanced design matrices as candidate regression models. A prerequisite for condition (7) to hold is

$$\zeta_n = \inf_{w \in Q_n} R(w) \rightarrow \infty, \quad (10)$$

which is especially relevant under the context of high-dimensional regression.

Condition (7) sets an upper bound on the number M of models for averaging. With (8), the condition (7) holds if

$$M^{4K+2} n^K / \zeta_n^{2K} \rightarrow 0. \quad (11)$$

Now suppose that the order of ζ_n is $n^{1-\delta}$ with $\delta \geq 0$. Then condition (11) is reduced to $M^{1+0.5K^{-1}} / n^{(1-2\delta)/4} \rightarrow 0$. Because K is fixed and the term $0.5K^{-1}$ is ignorable for normally distributed errors, M is allowed to grow to infinity if $\delta < 1/2$.

The condition (7) imposes certain limitations on the situation to apply our asymptotic results. We want to allow for the number of models M to increase to infinity as the sample size increases. For this to happen, ζ_n should grow at a rate no slower than $n^{1/2}$.

For nonparametric regression application, the seminal article of Stone (1982) has established an universal optimal rate of convergence under various smoothness conditions. For example, when the true regression function is twice differentiable, the optimal rate for the root mean squared error (MSE) is $n^{-2/(4+d)}$, where d is the dimension of regressors. Translating this result to the total squared error loss used in our setting, the optimal rate becomes $n \times (n^{-2/(4+d)})^2 = n^{1-1/(1+d/4)} = n^{1-\delta}$. We see that as d increases, δ gets smaller; for $d > 4$, we have $\delta < 1/2$. Thus, our condition (7) is reasonable under the high-dimensional nonparametric regression setting.

In the general setting of regression with p covariates where p increases as n increases (especially $p > n$), it is difficult to find a prevailing optimal rate of convergence that can serve as an

universal lower bound. In fact, no results have yet been available that give clean-cut rates as simple as those in the nonparametric regression settings. The lasso-related articles imposed sparsity conditions, mainly for the purpose of guaranteeing their favorite estimators' ability to detect all regressors with nonzero regression coefficients. The asymptotic settings typically assume that the size of the smallest nonzero regression coefficient has to be large enough for nearly unambiguous detection (as implied by condition (2.9) of Zhang 2010, for example, also condition (H) of Fan and Peng 2004). Under such model, the oracle rate for prediction would be the same as the number d_0 of the nonzero parameters. However, the achievability of these rates comes with a price because the tuning (penalty) parameter must be set properly and the parameter range determination needs the magic of oracle.

Such a dichotomy assumption, either very large or very small (zero) effects of regression and nothing in-between, leaves a lot of room for discussion in application. An often-encountered challenge in gene expression study (e.g., using gene expression to predict human height or complex disease traits) is when a substantial number of nonzero regression coefficients (gene effects) are only modestly detectable (size $1/\sqrt{n}$). But when biologists expand the study by doubling the sample sizes, not only a good portion of the modestly significant genes may become highly significant but also more new genes may emerge as being modestly significant. This phenomenon renders the commonly assumed sparsity condition (as specified in most literature) vulnerable. While it is likely to have a handful of dominant genes but their effects may only explain a portion of total genetic effect. The remaining portion may be attributed to abundant modest or small gene effects. For such situations, it seems prudent to assume that the best possible (oracle) rate is no better than the rate encountered in high-dimensional nonparametric regression. Therefore, our condition (7) is reasonable under the general high-dimensional regression context.

Proof of Theorem 1. We need the conclusion of Lemma 3.1, which is stated after the end of proof. To save notation, we sometimes will liberally share the bounding constants (C , C' , etc.), when deriving inequalities. Let $\tilde{L}(\mathbf{w}) = (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})'(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})$ with $\tilde{\boldsymbol{\mu}} = \sum_{k=1}^M w_k \tilde{H}_k \mathbf{y}$. Let $\Sigma = \text{cov}(\boldsymbol{\epsilon})$, which is a diagonal matrix. We have

$$\begin{aligned} \text{CV}(\mathbf{w}) &= \|\boldsymbol{\epsilon}\|^2 + \tilde{L}(\mathbf{w}) + 2\langle \boldsymbol{\epsilon}, \boldsymbol{\mu} - \tilde{H}(\mathbf{w})\mathbf{y} \rangle \\ &= \|\boldsymbol{\epsilon}\|^2 + L(\mathbf{w}) \left(\frac{\tilde{L}(\mathbf{w})}{L(\mathbf{w})} + \frac{2\langle \boldsymbol{\epsilon}, \boldsymbol{\mu} - \tilde{H}(\mathbf{w})\mathbf{y} \rangle / R(\mathbf{w})}{L(\mathbf{w}) / R(\mathbf{w})} \right). \end{aligned}$$

Because $\hat{\mathbf{w}}$ minimizes $\text{CV}(\mathbf{w})$ over $\mathbf{w} \in \mathcal{Q}_n$, it also minimizes $\text{CV}(\mathbf{w}) - \|\boldsymbol{\epsilon}\|^2$ over $\mathbf{w} \in \mathcal{Q}_n$. The claim $L(\hat{\mathbf{w}}) / \inf_{\mathbf{w} \in \mathcal{Q}_n} L(\mathbf{w}) \rightarrow 1$ is valid if

$$\sup_{\mathbf{w} \in \mathcal{Q}_n} |\tilde{L}(\mathbf{w}) / L(\mathbf{w}) - 1| \rightarrow 0, \quad (12)$$

$$\sup_{\mathbf{w} \in \mathcal{Q}_n} \langle \boldsymbol{\epsilon}, \boldsymbol{\mu} - \tilde{H}(\mathbf{w})\mathbf{y} \rangle / \zeta_n \rightarrow 0, \quad (13)$$

$$\sup_{\mathbf{w} \in \mathcal{Q}_n} |L(\mathbf{w}) / R(\mathbf{w}) - 1| \rightarrow 0 \quad (14)$$

hold. Because of Cauchy-Schwartz inequality, we can bound

$$|\tilde{L}(\mathbf{w}) - L(\mathbf{w})| = | \langle (\tilde{H}(\mathbf{w}) - H(\mathbf{w}))\mathbf{y}, (\tilde{H}(\mathbf{w}) - H(\mathbf{w}))\mathbf{y} \rangle |$$

by $\|(\tilde{H}(\mathbf{w}) - H(\mathbf{w}))\mathbf{y}\|^2 + 2\sqrt{L(\mathbf{w})} \|(\tilde{H}(\mathbf{w}) - H(\mathbf{w}))\mathbf{y}\|$. Therefore to show (12), it is sufficient to prove

$$\sup_{\mathbf{w} \in \mathcal{Q}_n} \|(\tilde{H}(\mathbf{w}) - H(\mathbf{w}))\mathbf{y}\|^2 / L(\mathbf{w}) \rightarrow 0.$$

Now by triangle inequality, we can bound $\|(\tilde{H}(\mathbf{w}) - H(\mathbf{w}))\mathbf{y}\|^2$ by

$$\begin{aligned} & \left(\sum_{k=1}^M w_k \|(\tilde{H}_k - H_k)\boldsymbol{\mu}\| + \sum_{k=1}^M w_k \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\| \right)^2 \\ & \leq \left(\sum_{k=1}^M \|(\tilde{H}_k - H_k)\boldsymbol{\mu}\| + \sum_{k=1}^M \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\| \right)^2 \\ & \leq M^2 \left(\max_{k=1, \dots, M} \|(\tilde{H}_k - H_k)\boldsymbol{\mu}\| + \max_{k=1, \dots, M} \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\| \right)^2 \\ & \leq 2M^2 \left(\max_{k=1, \dots, M} \|(\tilde{H}_k - H_k)\boldsymbol{\mu}\|^2 + \max_{k=1, \dots, M} \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\|^2 \right). \end{aligned}$$

Therefore, by (14), it is sufficient to show

$$M^2 \max_{k=1, \dots, M} \|(\tilde{H}_k - H_k)\boldsymbol{\mu}\|^2 / \zeta_n \rightarrow 0, \quad (15)$$

$$M^2 \max_{k=1, \dots, M} \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\|^2 / \zeta_n \rightarrow 0. \quad (16)$$

Using Lemma 3.1, we have

$$\begin{aligned} & \|(\tilde{H}_k - H_k)\boldsymbol{\mu}\|^2 / \zeta_n \\ & \leq (\lambda_{\max}(\tilde{H}_k - H_k))^2 \|\boldsymbol{\mu}\|^2 / \zeta_n \\ & \leq C^2 \cdot \left(\frac{p_k}{n} \right)^2 \cdot \|\boldsymbol{\mu}\|^2 / \zeta_n \quad (\text{by (19) of Lemma 3.1}) \\ & \leq (\Lambda')^2 \cdot C_2 \cdot C^2 \cdot \sqrt{n} / \zeta_n \quad (\text{by conditions (6) and (8)}), \end{aligned}$$

where $\lambda_{\max}(\cdot)$ denotes the maximum singular value of a matrix. Therefore, Equation (15) follows from condition (7). To prove (16), it is sufficient to show

$$M^2 \max_{k=1, \dots, M} E \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\|^2 / \zeta_n \rightarrow 0 \quad (17)$$

and for any $\delta > 0$,

$$\sum_{k=1}^M P(M^2 \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\|^2 - E \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\|^2 / \zeta_n > \delta) \rightarrow 0. \quad (18)$$

Now

$$\begin{aligned} E \|(\tilde{H}_k - H_k)\boldsymbol{\epsilon}\|^2 &= \text{tr}(\tilde{H}_k - H_k) \Sigma (\tilde{H}_k - H_k)' \\ &\leq \lambda_{\max}(\Sigma) \cdot \text{tr}(\tilde{H}_k - H_k)(\tilde{H}_k - H_k)' \\ &\leq \lambda_{\max}(\Sigma) \cdot C^2 \cdot \frac{p_k^2}{n} \\ &\quad (\text{by (20) of Lemma 3.1}). \end{aligned}$$

Therefore, (17) holds. We can use Whittle (1960) to prove (18) in the following way:

$$\begin{aligned}
& \sum_{k=1}^M P(M^{4K} |||(\tilde{H}_k - H_k)\boldsymbol{\varepsilon}||^2 - E||(\tilde{H}_k - H_k)\boldsymbol{\varepsilon}||^2 / \zeta_n^{2K} > \delta^{2K}) \\
& \leq \sum_{k=1}^M \frac{M^{4K} E[||(\tilde{H}_k - H_k)\boldsymbol{\varepsilon}||^2 - E||(\tilde{H}_k - H_k)\boldsymbol{\varepsilon}||^2]^{2K}}{\zeta_n^{2K} \delta^{2K}} \\
& \leq C' \sum_{k=1}^M \frac{M^{4K}}{\zeta_n^{2K} \delta^{2K}} [\text{tr}((\tilde{H}_k - H_k)'(\tilde{H}_k - H_k))^2]^K \\
& \leq C' \cdot C \cdot \frac{M^{4K}}{\zeta_n^{2K} \delta^{2K}} \sum_{k=1}^M \left(\frac{p_k^4}{n^3}\right)^K \quad (\text{by (21) of Lemma 3.1}) \\
& \leq C' \cdot C \cdot (\Lambda')^K \cdot \frac{M^{4K+1}}{\zeta_n^{2K} \delta^{2K}} \rightarrow 0,
\end{aligned}$$

where C' is some constant. Thus, (18) is proved. Next, we prove (13). First, because

$$\begin{aligned}
& |\langle \boldsymbol{\varepsilon}, \boldsymbol{\mu} - \tilde{H}(\mathbf{w})\mathbf{y} \rangle| \\
& \leq |\langle \boldsymbol{\varepsilon}, \boldsymbol{\mu} \rangle| + \sum_{k=1}^M w_k |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\mu} \rangle| + \sum_{k=1}^M w_k |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\varepsilon} \rangle| \\
& \leq |\langle \boldsymbol{\varepsilon}, \boldsymbol{\mu} \rangle| + \sum_{k=1}^M |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\mu} \rangle| + \sum_{k=1}^M |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\varepsilon} \rangle| \\
& \leq |\langle \boldsymbol{\varepsilon}, \boldsymbol{\mu} \rangle| + M \max_{1 \leq k \leq M} |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\mu} \rangle| + M \max_{1 \leq k \leq M} |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\varepsilon} \rangle|,
\end{aligned}$$

it is sufficient to show that $|\langle \boldsymbol{\varepsilon}, \boldsymbol{\mu} \rangle|/\zeta_n \rightarrow 0$ (which is obvious) and for any $\delta > 0$,

$$\begin{aligned}
& \sum_{k=1}^M P(M|\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\mu} \rangle| \zeta^{-1} > \delta) \rightarrow 0, \\
& \sum_{k=1}^M P(M|\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\varepsilon} \rangle| \zeta^{-1} > \delta) \rightarrow 0.
\end{aligned}$$

Now,

$$\begin{aligned}
& \sum_{k=1}^M P(M|\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\mu} \rangle| \zeta^{-1} > \delta) \\
& \leq \sum_{k=1}^M P(M^{2K} |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\mu} \rangle|^{2K} \zeta^{-2K} > \delta^{2K}) \\
& \leq \frac{M^{2K}}{\zeta_n^{2K} \delta^{2K}} \sum_{k=1}^M E|\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\mu} \rangle|^{2K} \\
& \leq C \cdot \frac{M^{2K}}{\zeta_n^{2K} \delta^{2K}} \sum_{k=1}^M ||\tilde{H}_k \boldsymbol{\mu}||^{2K} \\
& \leq C \cdot \frac{M^{2K}}{\zeta_n^{2K} \delta^{2K}} \sum_{k=1}^M \lambda_{\max}(\tilde{H}_k)^{2K} ||\boldsymbol{\mu}||^{2K} \\
& \leq C \cdot (1 + C)^{2K} \cdot \frac{M^{2K+1}}{\zeta_n^{2K} \delta^{2K}} \cdot ||\boldsymbol{\mu}||^{2K} \rightarrow 0, \\
& \quad (\text{by (22) of Lemma 3.1}),
\end{aligned}$$

where C is some constant. Similarly,

$$\begin{aligned}
& \sum_{k=1}^M P(M|\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\varepsilon} \rangle| \zeta^{-1} > \delta) \\
& \leq \sum_{k=1}^M P(M^{2K} |\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\varepsilon} \rangle|^{2K} \zeta^{-2K} > \delta^{2K}) \\
& \leq \frac{M^{2K}}{\zeta_n^{2K} \delta^{2K}} \sum_{k=1}^M E|\langle \boldsymbol{\varepsilon}, \tilde{H}_k \boldsymbol{\varepsilon} \rangle|^{2K} \\
& \leq C \cdot \frac{M^{2K}}{\zeta_n^{2K} \delta^{2K}} \sum_{k=1}^M (\text{tr}\{\tilde{H}_k \tilde{H}_k'\})^K \\
& \leq C \cdot C^K \cdot \frac{M^{2K+1}}{\zeta_n^{2K} \delta^{2K}} \cdot n^K \rightarrow 0.
\end{aligned}$$

We have proved (13). The proof of (14) is in the Appendix. The proof of Theorem 1 is completed.

We put together some inequalities concerning $\tilde{H}(\mathbf{w})$ and $H(\mathbf{w})$ in the following lemma.

Lemma 3.1. Under condition (5), there exists a constant $C > 0$ such that

$$\lambda_{\max}(\tilde{H}_k - H_k) \leq \frac{\bar{\lambda}(H_k)}{1 - \bar{\lambda}(H_k)} \leq C \times \frac{p_k}{n} \quad (19)$$

$$\text{tr}(\tilde{H}_k - H_k)'(\tilde{H}_k - H_k) = \left(\frac{\bar{\lambda}(H_k)}{1 - \bar{\lambda}(H_k)}\right)^2 (n - p_k) \leq C^2 \times \frac{p_k^2}{n} \quad (20)$$

$$\begin{aligned}
& \text{tr}((\tilde{H}_k - H_k)'(\tilde{H}_k - H_k))^2 \\
& \leq \{\lambda_{\max}(\tilde{H}_k - H_k)\}^2 \cdot \text{tr}(\tilde{H}_k - H_k)'(\tilde{H}_k - H_k) \\
& \leq \left(\frac{\bar{\lambda}(H_k)}{1 - \bar{\lambda}(H_k)}\right)^4 (n - p_k) \\
& \leq C \cdot \frac{p_k^4}{n^3} \quad (21)
\end{aligned}$$

$$\lambda_{\max}(\tilde{H}_k) \leq \lambda_{\max}(H_k) + \lambda_{\max}(\tilde{H}_k - H_k) \leq 1 + C \times \frac{p_k}{n} \quad (22)$$

$$\text{tr}\{\tilde{H}_k \tilde{H}_k'\} \leq \left(\frac{1}{1 - \bar{\lambda}(H_k)}\right)^2 p_k \leq C \times p_k. \quad (23)$$

The proof of Lemma 3.1 is based on the relationship $\tilde{H}_k - H_k = S_k(I - H_k)$, where S_k is a diagonal matrix with i th element equal to $h_{k,ii}/(1 - h_{k,ii})$. Here $h_{k,ii}$ is the i th element of H_k . We also need to use the subadditivity and submultiplicability properties of $\lambda_{\max}(\cdot)$, $\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$, $\lambda_{\max}(AB) \leq \lambda_{\max}(A) \cdot \lambda_{\max}(B)$, and the inequality $\text{tr}(ABA') \leq \lambda_{\max}(B) \cdot \text{tr}\{AA'\}$.

3.2 Ordering of Regressors for Grouping

Marginal correlations between the predictors and the response have long been used for variable screening; see, for example, Fan and Lv (2008). In a situation in which no information is available except for a set of n observations, we calculate these correlations to group the variables with similar correlations together. For some applications such as gene expression studies,

genes with similar marginal correlations with the response variable may be indicative of sharing relevant biological functions. Recall that our procedure proposed at Step 1 discarded the last group of regressors with correlations closest to 0. To obtain better prediction, ideally we hope that marginal correlation can separate the informative regressors from the noninformative ones. Because we are most interested in the case $p > n$, an important issue concerns the minimum sample size n required. Lemma 3.1 suggests that $n \gg \log p$ may be unavoidable.

By rearranging β_j and without loss of generality, we partition each of the “observed predictors into signal and nonsignal components, $X = (X_T, X_F)$, where X_T and X_F are $n \times s$ and $n \times (p - s)$ submatrices, respectively. X_T contains the set of predictors to be included in the model.

Assuming that the design matrix X has been standardized (i.e., for each j , the j th column of X satisfies $n^{-1} \|X_{(j)}\|^2 = n^{-1} X'_{(j)} X_{(j)} = 1$), we estimate the marginal regression coefficients as

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{\mathbf{y}}_T \\ \hat{\mathbf{y}}_F \end{pmatrix} = \frac{1}{n} X' \mathbf{y} = \frac{1}{n} \begin{pmatrix} X'_T \mathbf{y} \\ X'_F \mathbf{y} \end{pmatrix}.$$

If there is no noise in \mathbf{y} , then the marginal regression coefficients are

$$\mathbf{y}_T = \frac{1}{n} X'_T (X_T \boldsymbol{\beta} + \boldsymbol{\varepsilon})|_{\boldsymbol{\varepsilon}=\mathbf{0}} = \frac{1}{n} X'_T X_T \boldsymbol{\beta}$$

and

$$\mathbf{y}_F = \frac{1}{n} X'_F (X_T \boldsymbol{\beta} + \boldsymbol{\varepsilon})|_{\boldsymbol{\varepsilon}=\mathbf{0}} = \frac{1}{n} X'_F X_T \boldsymbol{\beta}.$$

In order for informative regressors to be fully retained by marginal correlation under the noise-free case, we need the condition that $|\gamma_j| > k$ when $\beta_j \neq 0$ and $|\gamma_j| \leq k$ when $\beta_j = 0$. This assumption is the property of “faithfulness” coined by Buhlmann, Kalisch, and Maathuis (2010).

Under the usual signal plus noise case, the ordering of the observed marginal correlations is subject to the perturbation of noises. We investigate the sample size needed to overcome the influence of noise.

Lemma 3.2. Under the assumptions

$$\frac{1}{n} \min_{j \in T} |X'_{(j)} X_T \boldsymbol{\beta}| - \frac{1}{n} \max_{j \in F} |X'_{(j)} X_T \boldsymbol{\beta}| > 0$$

and $\log(p) = o(n)$, we have

$$P \left(\max_{j \in F} |\hat{\mathbf{y}}_F| > \min_{j \in T} |\hat{\mathbf{y}}_T| \right) \rightarrow 0,$$

where $F = \{s + 1, \dots, p\}$ and $T = \{1, \dots, s\}$ denote the false and true index sets, respectively.

Thus, we need $\log(p) = o(n)$, which implies that the order of p is smaller than the exponential of sample size n .

Proof of Lemma 3.2. We evaluate the following probability:

$$P \left(\max_{j \in F} |\hat{\mathbf{y}}_F| > \min_{j \in T} |\hat{\mathbf{y}}_T| \right),$$

where $F = \{s + 1, \dots, p\}$ and $T = \{1, \dots, s\}$ denote the false and true index sets, respectively, the operator \max takes the maximum value of one (or some) of the element(s) of $\hat{\mathbf{y}}_F$ and

\min takes the minimum value of one (or some) of the element(s) of $\hat{\mathbf{y}}_T$:

$$\begin{aligned} & P \left(\max_{j \in F} |\hat{\mathbf{y}}_F| > \min_{j \in T} |\hat{\mathbf{y}}_T| \right) \\ & \leq P \left(\max_{j \in F} |X'_{(j)} X_T \boldsymbol{\beta}| + \max_{j \in F} |X'_{(j)} \boldsymbol{\varepsilon}| > \min_{j \in T} |X'_{(j)} X_T \boldsymbol{\beta}| \right. \\ & \quad \left. - \max_{j \in T} |X'_{(j)} \boldsymbol{\varepsilon}| \right) \\ & = P \left(\max_{j \in T} |X'_{(j)} \boldsymbol{\varepsilon}| + \max_{j \in F} |X'_{(j)} \boldsymbol{\varepsilon}| > \min_{j \in T} |X'_{(j)} X_T \boldsymbol{\beta}| \right. \\ & \quad \left. - \max_{j \in F} |X'_{(j)} X_T \boldsymbol{\beta}| \right) \\ & \leq \frac{E \left[\max_{j \in T} |X'_{(j)} \boldsymbol{\varepsilon}| + \max_{j \in F} |X'_{(j)} \boldsymbol{\varepsilon}| \right]}{\min_{j \in T} |X'_{(j)} X_T \boldsymbol{\beta}| - \max_{j \in F} |X'_{(j)} X_T \boldsymbol{\beta}|} \\ & \quad \text{(Markov inequality).} \end{aligned}$$

For any Gaussian random vector (X_1, \dots, X_q) , we have

$$E \left[\max_j X_{(j)} \right] \leq 3\sqrt{\log q} \times \max_j \sqrt{E[X_{(j)}^2]}.$$

Thus, noting that $V(X'_{(j)} \boldsymbol{\varepsilon}) = X'_{(j)} V(\boldsymbol{\varepsilon}) X_{(j)} = \sigma^2 \|X_{(j)}\|^2$, we have

$$\begin{aligned} E \left[\max_{j \in T} |X'_{(j)} \boldsymbol{\varepsilon}| \right] & \leq 3\sqrt{\log s} \times \max_{j \in T} \sqrt{V(X'_{(j)} \boldsymbol{\varepsilon})} \\ & \leq 3\sqrt{\log s} \times \sigma \max_{j \in T} \|X_{(j)}\| \end{aligned}$$

and

$$\begin{aligned} E \left[\max_{j \in F} |X'_{(j)} \boldsymbol{\varepsilon}| \right] & \leq 3\sqrt{\log(p-s)} \times \max_{j \in F} \sqrt{V(X'_{(j)} \boldsymbol{\varepsilon})} \\ & \leq 3\sqrt{\log(p-s)} \times \sigma \max_{j \in F} \|X_{(j)}\|. \end{aligned}$$

We also have

$$\begin{aligned} & P \left(\max_{j \in F} |\hat{\mathbf{y}}_F| > \min_{j \in T} |\hat{\mathbf{y}}_T| \right) \\ & \leq \left(3\sqrt{\log s} \times \sigma \max_{j \in T} \|X_{(j)}\| + 3\sqrt{\log(p-s)} \right. \\ & \quad \left. \times \sigma \max_{j \in F} \|X_{(j)}\| \right) / \left(\min_{j \in T} |X'_{(j)} X_T \boldsymbol{\beta}| - \max_{j \in F} |X'_{(j)} X_T \boldsymbol{\beta}| \right) \\ & = \left(3\sigma \left(\sqrt{\log s} \max_{j \in T} \|X_{(j)}\| / \sqrt{n} + \sqrt{\log(p-s)} \right. \right. \\ & \quad \left. \left. \times \max_{j \in F} \|X_{(j)}\| / \sqrt{n} \right) \right) \\ & \quad / \left(\sqrt{n} \times \left(n^{-1} \min_{j \in T} |X'_{(j)} X_T \boldsymbol{\beta}| - n^{-1} \max_{j \in F} |X'_{(j)} X_T \boldsymbol{\beta}| \right) \right). \end{aligned}$$

We now investigate the condition that leads to

$$P \left(\max_{j \in F} |\hat{\mathbf{y}}_F| > \min_{j \in T} |\hat{\mathbf{y}}_T| \right) \rightarrow 0 \quad n \rightarrow \infty.$$

Noting that $\max_{j \in T} \|X_{(j)}\| / \sqrt{n} = 1$ and $\max_{j \in F} \|X_{(j)}\| / \sqrt{n} = 1$, we require

$$\begin{aligned} & \frac{3\sigma(\sqrt{\log s} + \sqrt{\log(p-s)})}{\sqrt{n} \times (n^{-1} \min_{j \in T} |X'_{(j)} X_T \boldsymbol{\beta}| - n^{-1} \max_{j \in F} |X'_{(j)} X_T \boldsymbol{\beta}|)} \\ & = o(1), \end{aligned}$$

which implies

$$3 \times 2\sigma \sqrt{\frac{\log p}{n}} < \left(\frac{1}{n} \min_{j \in T} |X'_{(j)} X_T \beta| - \frac{1}{n} \max_{j \in F} |X'_{(j)} X_T \beta| \right).$$

Here, we use the fact that p is much greater than the number of true predictors s . We immediately see that we require

$$\frac{1}{n} \min_{j \in T} |X'_{(j)} X_T \beta| - \frac{1}{n} \max_{j \in F} |X'_{(j)} X_T \beta| > 0.$$

Otherwise, it is impossible to reconstruct the true structure based on the marginal regression. This completes the proof. \square

3.3 Optimal Grouping for Risk Minimization

In general, the comparison of model averaging between different classes of models is difficult to study. The pattern of signals and the covariance structure of regressors are important factors to consider. To gain some insight, we consider a simplified situation wherein an analytic solution is available for $\min_{Q_n} R(w)$.

Suppose there are K regressors that have nonzero marginal correlations with the outcome variable. Suppose we want to form two models by dividing them into two groups of size K_1 , K_2 , respectively ($K_1 + K_2 = K$). Let $m_1 = K_1/K$ and $m_2 = K_2/K$. Assume that these regressors are orthogonal to each other. Assume further that each regressor has variance equal to 1 and the noise variance is equal to σ^2 .

Then a simple calculation shows that the risk

$$R(w_1, w_2) = E(L(w_1, w_2)) = \sum_{i=1,2} (1 - w_i)^2 (\|\mu_i\|^2) + w_i^2 K_i \sigma^2,$$

where μ_1 and μ_2 are the true mean vector μ projected on the space spanned by the vectors of regressors in model 1 and in model 2, respectively, $\mu = \mu_1 + \mu_2$.

Let β_j be true regression parameter for regressor j . Then for $i = 1, 2$, $\|\mu_i\|^2 = n \sum_{j \in \text{model}_i} \beta_j^2$. Straightforward calculation shows that

$$\begin{aligned} \min_{0 \leq w_1, w_2 \leq 1} R(w_1, w_2) &= \sum_{i=1,2} \frac{\|\mu_i\|^2 K_i \sigma^2}{\|\mu_i\|^2 + K_i \sigma^2} \\ &= K \sigma^2 \sum_{i=1,2} m_i \left(1 - \frac{\sigma^2}{\|\mu_i\|^2 / K_i + \sigma^2} \right) \\ &= K \sigma^2 (1 - E(\sigma^2 / Z + \sigma^2)), \end{aligned}$$

where Z denotes the random variable taking the value of $\|\mu_i\|^2 / K_i$, with probability m_i , $i = 1, 2$.

Note that the graph of the function $1/(z + \sigma^2)$ is convex in z . Geometrically, it is easy to see that for any four points $z_1 \geq z_3 \geq z_4 \geq z_2$, the line segment connecting z_1, z_2 lies above the line segment connecting z_3 and z_4 . This allows us to show that the partition by ordering the size of correlation minimizes the risk.

Lemma 3.3. Among all possible partitions with two groups of size K_1 and K_2 , $E(\sigma^2 / Z + \sigma^2)$ is maximized when the K_1 largest values of β_j^2 belong to one group and the rest of them belong to another group.

The same argument can be applied to more than two groups. Suppose M groups are to be created with fixed sizes,

K_1, \dots, K_M . We can apply the same argument of Lemma 3.3 iteratively to show that partition after ordering the correlation by size is an optimal way of generating K models for model averaging.

4. NUMERICAL RESULTS

There are two main purposes in this section: (i) to demonstrate the gain in relaxing the constraint of weights summing to 1 in model averaging and (ii) to compare the performance of model averaging with various shrinkage methods of high-dimensional regression. Because the performance of model averaging may depend on the class of models used for averaging, two versions of implementing the proposed procedure in Step 1 of Section 2 are considered.

The first version, MCV1, follows Step 1 to order regressors by correlation with output variables. The number of regressors p_k in each model is set to be the same, $p_k = nh$, where h is a value between 0 and 1. The value of h and the value of M (the total number of models) will be specified later on.

Because the specification of M and h may vary case by case, here we propose a practical strategy for optimizing the choice of M and h . First, we determine the number T of potentially informative regressors. In our implementation, we set T to be the number of regressors whose marginal correlations with the output variable are deemed as statistically significant at the 5% level. Because the total number of regressors from M models is equal to Mhn , it is desirable for Mhn to be no greater than T . Subject to this restriction, we choose M and h whose cross-validation prediction error is as small as possible. The version MCV2 implements this strategy using a grid search of optimal h . Note that it is possible that some of the T significant regressors are precluded in the class of models given an allowable choice of M and h . To compensate for this omission, instead of requiring that all models to have an equal number of regressors, MCV2 allows the number of regressors in the last model of a class to be between 1 and hn .

4.1 Simulation Study

In this simulation study, we adopted the linear model used in (11). Following the settings of Buhlmann, Kalisch, and Maathuis (2010), we generate the possible p predictors from the normal with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$ with $s_{ij} = \rho^{|i-j|}$. The true coefficients β_j are generated from the normal with mean of 0 and standard deviation (SD) of 0.5. Seven model settings are considered:

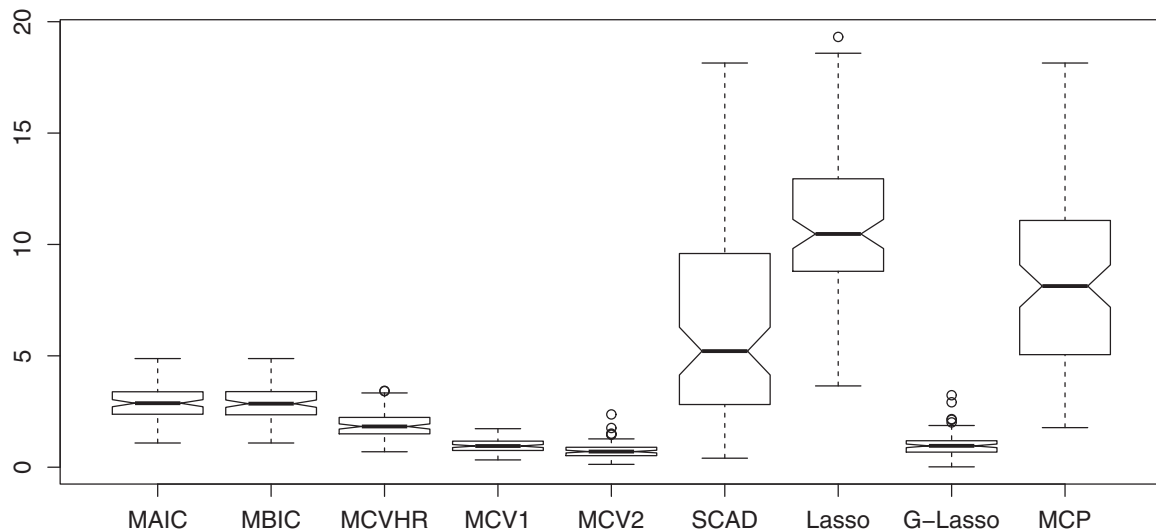
- Set sample size $n = 50$ and the number of regressors $p = 2000$. Set the number of true regressors $s = 50$ and let the true regressors x_i be spaced evenly, $i = 40(j-1) + 1$, $j = 1, \dots, 50$. The value of ρ is set to 0.6. Normal noise with mean of 0 and SD of 0.2 is generated.
- Following a comment from referee, we consider the following setting. We set sample size $n = 50$, the number of regressors $p = 2000$, the number of true regressors $s = 50$, and let the true regressors x_i be spaced evenly, $i = 40(j-1) + 1$, $j = 1, \dots, 50$. The value of ρ is set to 0.5. Noise are generated

from a mixture of normal and Student's- t distribution: $\varepsilon_\alpha \sim s_\alpha \times N(0, 1)$ with the probability 0.5, and $\varepsilon_\alpha \sim s_\alpha \times \text{St}(0, \nu = 5)$ with the probability 0.5, for $\alpha = 1, \dots, n$. Here $s_\alpha = \sqrt{\sum_{i=1}^{50} |x_i|/50}$ is a scaling factor, which depends on the magnitude of the true regressors. Also, the true coefficients β_j are generated from a mixture of normal and Student's- t distribution: $\varepsilon_\alpha \sim s_\alpha \times N(0, 0.8)$ with the probability 0.7, and $\varepsilon_\alpha \sim s_\alpha \times \text{St}(0, \nu = 3)$ with the probability

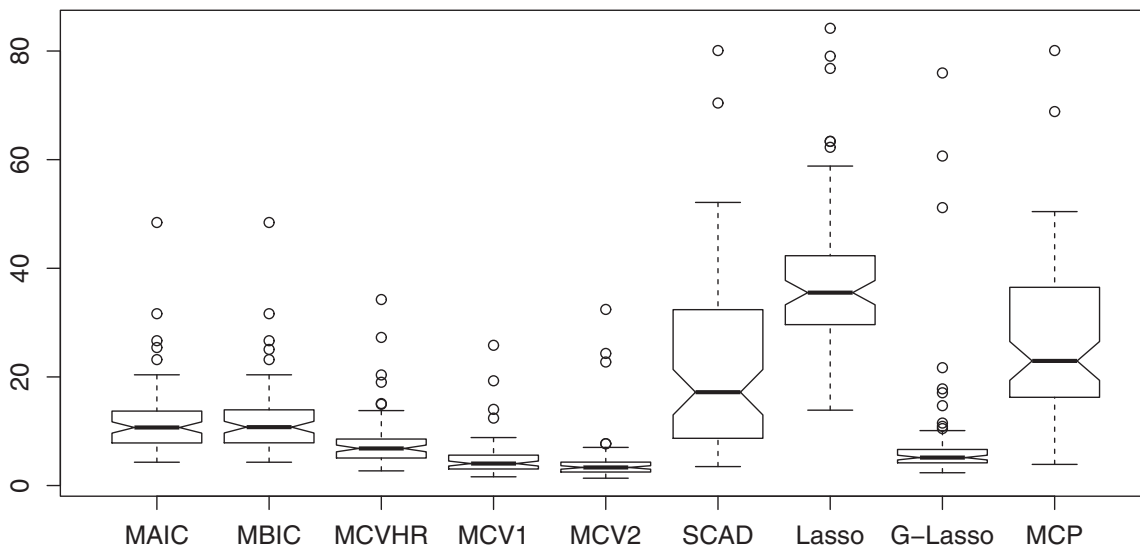
0.3. This setting creates more heterogeneity than the setting in (a).

(c)–(g) See the online supplementary materials.

For MCV1, we set $M = 10$ and $nh = 10$ to yield a class of 10 models, each with 10 regressors. This class of models is also used in implementing three methods of weight selection under the traditional model averaging, which requires the constraint of total weights summing to one: MCVHR (the delete-one



Simulation design (a).



Simulation design (b).

Figure 1. Boxplots of the performance measure MSE of the simulation designs (a) and (b). MAIC, model averaging with the Akaike information criterion under the restriction $\sum_{k=1}^M w_k = 1$; MBIC, model averaging with the Bayesian information criterion under the restriction $\sum_{k=1}^M w_k = 1$; MCVHR, model averaging with cross-validation under the restriction $\sum_{k=1}^M w_k = 1$ (Hansen and Racine 2012); MCV1, cross-validation model averaging without the restriction $\sum_{k=1}^M w_k = 1$ and the number of models and the number of regressors in each model is fixed; MCV2, cross-validation model averaging without the restriction $\sum_{k=1}^M w_k = 1$ and the number of models M and the number of regressors in each model are optimized; SCAD, penalized regression by SCAD approach; Lasso, original lasso procedure; G-Lasso, group lasso procedure; MCP, panelized regression by MCP approach.

cross-validation of Hansen and Racine 2012), the Akaike information criterion (MAIC), and the Bayesian information criterion (MBIC).

The implementation of MCV2 is carried out by setting the grid points of h , $h = 0.05j$, $j = 1, \dots, 8$ for $n = 50$ and $h = 0.025j$, $j = 1, \dots, 8$ for $n = 100$. The value of M is free to vary. The maximum possible value of M is automatically determined by T , h , and n .

To implement the MCP (Zhang 2010; Breheny and Huang 2011) and SCAD (Fan and Li 2001) algorithms, the R package program `ncvreg` was used. 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.

The original lasso (Tibshirani 1996) and group lasso (Yuan and Lin 2006) are also considered for comparison. Lasso is implemented by `lars` package in R (see Efron et al. 2004). Since the desirable performance of lasso heavily hinge on an appropriate selection of the tuning parameter, which controls the tradeoff between the lack of fit and the size of regression coefficient parameters, we used the BIC (Wang, Li, and Leng 2009) for selection.

For implementing group lasso, we partitioned the regressors into $M + 1$ groups. The first M groups are the same as those obtained in MCV1, MCVHR, MAIC, and MBIC. The last group consists of all the remaining regressors. The R-package program `grplasso` was used. An optimal value of the tuning parameter can be found using a modified version of the C_p criterion. However, exact computation of the degrees of freedom is difficult. Although Yuan and Lin (2006) proposed the use of bootstrapping, this is a computationally intensive task. Moreover, the true

variance of the noise required for applying the C_p criterion is unknown. Thus, we used the $k(= 5)$ -fold cross-validation procedure. A set of candidate values of the regularization parameter was prepared by using the function `lamdamax` in the R package program `grplasso`.

Table 1 compares the computational time required for each method. We can see that the computational time required for our method is very small. Here, we repeat the simulation 100 times and the computing time required in each time is recorded. The table gives the mean (in seconds) and corresponding SD for each method. Since the required computational times for MAIC and MBIC are the same, we just report the computational time required for MAIC. The computation times under the settings (b)–(g) have the same trends as (a); the data are omitted in Table 1.

We used MSE (averaged squared difference between the true μ and the estimated μ) as the performance measure for each method. Figure 1 shows the boxplot of MSEs after 100 simulation runs under the settings (a) and (b) (figures under the remaining settings are provided as supplementary materials). As shown in these figures, our model-averaging procedure yields a nice performance in the sense that it achieves the smallest MSE median. The results show that it is beneficial to relax the weight constraint in model averaging. It also showed that the proposed model-averaging approach yields comparable or better performance than the lasso-type of methods.

4.2 Real Data Application

In this section, the proposed method is applied to the birth-weight dataset from Hosmer and Lemeshow (1989). The output variable contains the birthweights of $n = 189$ babies. Among

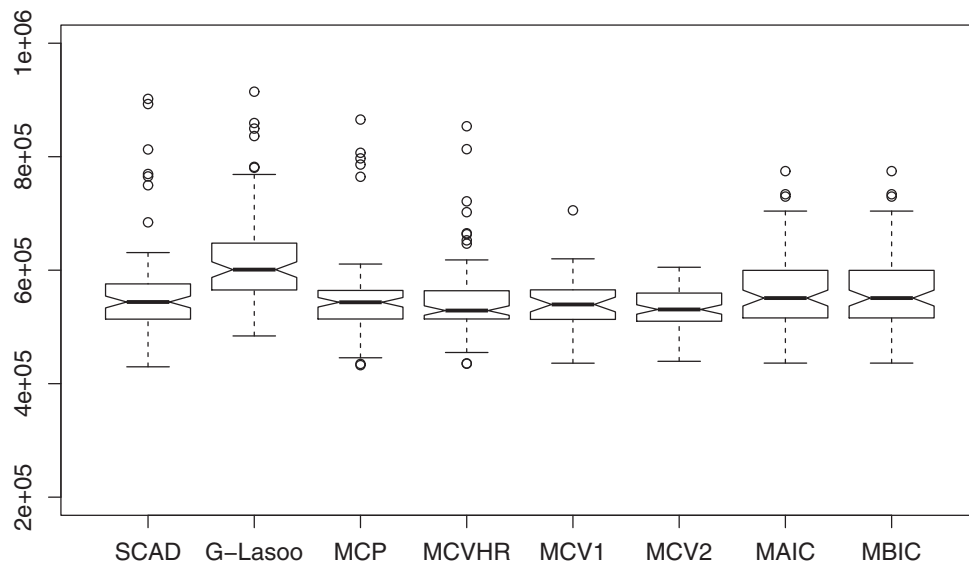


Figure 2. Birth weight data. Boxplots for the prediction errors of each method after 100 simulation runs. MAIC, model averaging with the Akaike information criterion under the restriction $\sum_{k=1}^M w_k = 1$; MBIC, model averaging with the Bayesian information criterion under the restriction $\sum_{k=1}^M w_k = 1$; MCVHR, model averaging with cross-validation under the restriction $\sum_{k=1}^M w_k = 1$ (Hansen and Racine 2012); MCV1, cross-validation model averaging without the restriction $\sum_{k=1}^M w_k = 1$ and the number of models and the number of regressors in each model is fixed; MCV2, cross-validation model averaging without the restriction $\sum_{k=1}^M w_k = 1$ and the number of models M and the number of regressors in each model are optimized; SCAD, penalized regression by SCAD approach; G-Lasso, group lasso procedure; MCP, panelized regression by MCP approach.

the eight predictors, mother's age in years (x_1) and mother's weight in pounds at last menstrual period (x_2) are continuous predictors. The remaining six variables (x_3, \dots, x_8) are categorical: mother's race (white, black, or other), smoking status during pregnancy (yes or no), number of previous premature labors (0, 1, or 2), history of hypertension (yes or no), presence of uterine irritability (yes or no), and number of physician visits during the first trimester (0, 1, 2, or 3).

We used the interactions between the above eight regressors of order 3 or less of the form $x_i x_j x_k$ as predictors. Some of these predictors have correlations higher than 0.99 with others. We deleted these predictors and ended up with a list of 84 predictors as given in the table in the supplementary materials. To capture nonlinear effects of the continuous variables x_1 and x_2 , we followed Yuan and Lin (2006) and added x_1^2 , x_1^3 , x_2^2 , and x_2^3 . Therefore, a set of $p = 88$ predictors will be used.

To evaluate the prediction performance of various methods, we randomly selected $n = 50$ observations for model fitting, and used the rest of the data as the test set. Note unlike Yuan and Li (2006) wherein three quarters of the n observations was used for model fitting, our choice of n is to simulate the situation of $p > n$. We repeated this process 100 times to obtain the distribution for the prediction errors.

For MCV1, MCVHR, MAIC, and MBIC, we set $M = 4$ and $nh = 5$. For MCV2, we set $h = j/50$ with $j = 4, 5, 6$ and let M to vary freely. For grouped lasso, we attempted to use the same partition method to generate groups as described earlier in the section of Simulation Study. We selected the optimal penalty based on five-fold cross-validation.

The prediction errors for the testing data were shown in Figure 2. We can see that our method has very good prediction performance.

5. CONCLUSION

Recently, the analysis of high-dimensional data has attracted a lot of attention. Most works were driven under the signal sparseness condition and the focus was on the signal detection and estimation. In this article, we addressed the companion issue of prediction accuracy and introduced a model-averaging procedure that can be applied in the case where the number of regressors is greater than the sample size. Our method used marginal correlation to group regressors for model averaging and applied delete-one cross-validation to select optimal weights without the conventional constraint of total weights summing to 1.

Conceptually, cross-validation allows the data to determine the appropriate degree of shrinkage of each individual model predictor before combining together. This is an advantage over existing lasso methods that have not resolved the difficulties encountered in the penalty parameter selection.

We are able to derive an optimality property of cross-validation under the condition that each individual model in the class for averaging has a prediction error much higher than the standard root n rate in parametric regression. We believe this condition is an appropriate description of the situation often encountered in the application of high-dimensional regression.

We demonstrated that our method performed favorably in comparison with other methods in a simulation study and a real data example. However, the complexity of high-dimension regression makes it difficult to find an universally optimal solution. The optimality of Theorem 3.1 suggests that the performance of model averaging depends on the class of models proposed for averaging. To determine which class of models for averaging will work the best may hinge on the space where the true mean vector μ may belong to. The covariance structure of the regressors further complicates the issue. This and many other unsettled questions deserve further studies.

APPENDIX A. VALIDITY OF CONDITION (14)

We shall prove claim (14), which is equivalent to

$$\sup_{w \in Q_n} \left| \frac{\|H(w)\epsilon\|^2 - \text{tr}(H(w)H(w)\Sigma) - 2\langle A(w)\mu, H(w)\epsilon \rangle}{R(w)} \right| \rightarrow 0,$$

where $A(w) = I - H(w)$. It suffices to show that

$$\sup_{w \in Q_n} \left| \frac{\|H(w)\epsilon\|^2 - \text{tr}(H(w)H(w)\Sigma)}{R(w)} \right| \rightarrow 0 \quad (\text{A.1})$$

and

$$\sup_{w \in Q_n} \left| \frac{\langle A(w)\mu, H(w)\epsilon \rangle}{R(w)} \right| \rightarrow 0. \quad (\text{A.2})$$

We first establish claim (A.1). For any $\delta > 0$, we have

$$\begin{aligned} & P \left(\sup_{w \in Q_n} \left| \frac{\|H(w)\epsilon\|^2 - \text{tr}(H(w)H(w)\Sigma)}{R(w)} \right| > \delta \right) \\ & \leq P \left(\sup_{w \in Q_n} \left| \|H(w)\epsilon\|^2 - \text{tr}(H(w)H(w)\Sigma) \right| > \delta \zeta_n \right) \\ & \leq P \left(\sup_{w \in Q_n} \sum_{k=1}^M \sum_{m=1}^M w_k w_m \left| \epsilon' H_k H_m \epsilon - \text{tr} H_k H_m \Sigma \right| > \delta \zeta_n \right) \\ & \leq P \left(M^2 \times \max_k \max_m \left| \epsilon' H_k H_m \epsilon - \text{tr} H_k H_m \Sigma \right| > \delta \zeta_n \right) \\ & \leq \sum_{k=1}^M \sum_{m=1}^M P \left(\left| \langle \epsilon' H_k, H_m \epsilon \rangle - \text{tr} H_k H_m \Sigma \right| > \delta \zeta_n / M^2 \right) \\ & \leq \sum_{k=1}^M \sum_{m=1}^M M^{4K} \delta^{-2K} \zeta_n^{-2K} E \left[\left| \langle \epsilon' H_k, H_m \epsilon \rangle - \text{tr} H_k H_m \Sigma \right|^{2K} \right] \\ & \leq C M^{4K} \delta^{-2K} \zeta_n^{-2K} \sum_{k=1}^M \sum_{m=1}^M \left\{ \text{tr} H_k^2 H_m^2 \right\}^K \\ & \leq C M^{4K} \delta^{-2K} \zeta_n^{-2K} \sum_{k=1}^M \sum_{m=1}^M \left\{ \text{tr} H_k H_m \right\}^K \\ & \leq C \delta^{-2K} \zeta_n^{-2K} M^{4K+2} n^K, \end{aligned}$$

where C is some constant. As established before, this term converges to zero. Similarly, for any $\delta > 0$,

$$\begin{aligned} & P \left(\sup_{w \in Q_n} \left| \frac{\langle A(w)\mu, H(w)\epsilon \rangle}{R(w)} \right| > \delta \right) \\ & \leq P \left(\sup_{w \in Q_n} \left| \sum_{k=1}^M \sum_{m=1}^M w_k w_m \mu' (I - H_k) H_m \epsilon \right| > \delta \zeta_n \right) \\ & \leq P \left(M^2 \max_{1 \leq k \leq M} \max_{1 \leq m \leq M} \left| \mu' (I - H_k) H_m \epsilon \right| > \delta \zeta_n \right) \\ & \leq \frac{M^{4K}}{\delta^{2K} \zeta_n^{2K}} \sum_{k=1}^M \sum_{m=1}^M E \left| \mu' (I - H_k) H_m \epsilon \right|^{2K} \\ & \leq C \cdot \frac{M^{4K}}{\delta^{2K} \zeta_n^{2K}} \sum_{k=1}^M \sum_{m=1}^M \|H_m (I - H_k) \mu\|^{2K} \end{aligned}$$

$$\begin{aligned} &\leq C \cdot \frac{M^{4K}}{\delta^{2K} \zeta_n^{2K}} \sum_{k=1}^M \sum_{m=1}^M \|(I - H_k)\mu\|^{2K} \\ &\leq C \cdot \frac{M^{4K+2}}{\delta^{2K} \zeta_n^{2K}} \cdot \|\mu\|^{2K}, \end{aligned}$$

where C is some constant. The last two inequalities are obtained by noting that both H_m and $(I - H_k)$ are the projection matrices. The last expression converges to zero because of condition (7). We thus proved the claim (A.2).

SUPPLEMENTARY MATERIALS

The supplementary materials include: (1) additional simulation results in Section 4.1, and (2) Table 2 (a list of 84 predictors used in Birth weight data) in Section 4.2.

[Received July 2012. Revised August 2013.]

REFERENCES

- Akaike, H. (1979), "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting," *Biometrika*, 66, 237–242. [254]
- Ando, T., and Tsay, R. (2010), "Predictive Likelihood for the Bayesian Model Selection and Averaging," *International Journal of Forecasting*, 26, 744–763. [254]
- Avramov, D. (2002), "Stock Return Predictability and Model Uncertainty," *Journal of Financial Economics*, 64, 423–458. [255]
- Breheny, P., and Huang, J. (2011), "Coordinate Descent Algorithms for Non-convex Penalized Regression, With Applications to Biological Feature selection," *Annals of Applied Statistics*, 5, 232–253. [254,263]
- 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. [254,260,261]
- Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is Much Larger than n ," *The Annals of Statistics*, 35, 2313–2351. [254]
- Crespo-Cuaresma, J., and Slacik, T. (2009), "On the Determinants of Currency Crises: The Role of Model Uncertainty," *Journal of Macroeconomics*, 31, 621–632. [255]
- Efron, B., Johnstone, I., Hastie, T., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [254,263]
- Eicher, T., Henn, C., and Papageorgiou, C. (2012), "Trade Creation and Diversion Revisited: Accounting for Model Uncertainty and Natural Trading Partner Effects," *Journal of Applied Econometrics*, 27, 296–321. [255]
- Eklund, J., and Karlsson, S. (2007), "Forecast Combination and Model Averaging Using Predictive Measures," *Econometric Reviews*, 26, 329–363. [254]
- Fan, J., and Li, R. (2001), "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [254,263]
- Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society, Series B*, 70, 849–911. [254,259]
- Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961. [258]
- Fernandez, C., Ley, E., and Steel, M. F. J. (2001), "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381–427. [254]
- Garratt, A., Lee, K., Pesaran, H., and Shin, Y. (2003), "Forecast Uncertainties in Macroeconomic Modeling: An Application to the U.K. Economy," *Journal of the American Statistical Association*, 98, 829–838. [255]
- Genovese, C., Jin, J., and Wasserman, L. (2009), "Revisiting Marginal Regression," available at <http://arxiv.org/abs/0911.4080v1> [254]
- Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [254]
- Hansen, B. E., and Racine, J. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [254,256,262,263]
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [254]
- Hosmer, D. W., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: Wiley. [263]
- Kass, R., and Raftery, A. (1995), "Bayes Factors and Model Uncertainty," *Journal of the American Statistical Association*, 90, 773–795. [254]
- Li, K.-C. (1986), "Asymptotic Optimality of C_L and Generalized Cross-validation in Ridge Regression With Application to Spline Smoothing," *The Annals of Statistics*, 14, 1011–1112. [255,256,257]
- (1987), "Asymptotic Optimality for C_p , C_L , Cross-validation and Generalized Crossvalidation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975. [255,257]
- Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546. [254]
- Mallows, C. L. (1973), "Some comments on C_p ," *Technometrics*, 15, 661–675. [254]
- Meier, L., van de Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society B*, 70, 53–71. [256]
- Morales, K., Ibrahim, J., Chen, C., and Ryan, L. (2006), "Bayesian Model Averaging With Applications to Benchmark Dose Estimation for Arsenic in Drinking Water," *Journal of the American Statistical Association*, 101, 9–17. [255]
- Newbold, P., and Granger, C. W. J. (1974), "Experience With Forecasting Univariate Time Series and the Combination of Forecasts" (with discussion), *Journal of the Royal Statistical Society, Series A*, 137, 131–149. [254]
- Palm, F. C., and Zellner, A. (1992), "To Combine or Not to Combine? Issues of Combining Forecasts," *Journal of Forecasting*, 11, 687–701. [254]
- Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [254]
- Pesaran, H., Schleicher, C., and Zaffaroni, P. (2009), "Model Averaging in Risk Management With an Application to Futures Markets," *Journal of Empirical Finance*, 16, 280–305. [255]
- Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191. [254]
- Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection" (with discussion), *Statistica Sinica*, 7, 221–264. [255]
- Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [257]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [254,263]
- Wan, A., and Zhang, X. (2009), "On the Use of Model Averaging in Tourism Research," *Annals of Tourism Research*, 36, 525–532. [255]
- Wan, A. T. K., Zhang, X., and Zou, G. (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283. [254]
- Wang, H., Li, B., and Leng, C. (2009), "Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters," *Journal of the Royal Statistical Society, Series B*, 71, 671–683. [263]
- Whittle, P. (1960), "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and its Applications*, 5, 302–305. [259]
- Wright, J. H. (2008), "Bayesian Model Averaging and Exchange Rate Forecasts," *Journal of Econometrics*, 146, 329–341. [255]
- Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67. [254,256,263,264]
- Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [258,263]
- Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [254]
- Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320. [254]
- Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the Degrees of Freedom of the Lasso," *The Annals of Statistics*, 35, 2173–2192. [254]