Combining Linear Regression Models: When and How?

Zheng Yuan and Yuhong Yang*

December, 2004

Abstract

Model combining (mixing) methods have been proposed in recent years to deal with uncertainty in model selection. Even though advantages of model combining over model selection have been demonstrated in simulations and data examples, it is still unclear to a large extent when model combining should be preferred. In this work, firstly, an instability measure to capture the uncertainty of model selection in estimation, named PIE, is proposed based on perturbation of the sample. It is demonstrated that estimators from model selection can have large PIE values and model combining substantially reduces the instability for such cases. Secondly, we propose a model combining method, ARMS, and derive a theoretical property. In ARMS, a screening step is taken to narrow down the list of candidate models before combining, which not only saves computing time but also can improve estimation accuracy. Thirdly, we compare ARMS with EBMA (an empirical Bayesian model averaging) and model selection methods in a number of simulations and real data examples. The comparison shows that model combining produces better estimators when the instability of model selection is high and ARMS performs better than EBMA in most such cases in our simulations. With respect to the choice between model selection and model combining, we propose a rule of thumb in terms of PIE. The empirical results support that PIE is a sensible indicator of model selection instability in estimation and is useful for understanding whether model combining is a better choice over model selection for the data at hand.

KEY WORDS: Adaptive regression by mixing; Bayesian model averaging; Model combining; Model selection; Model uncertainty; Instability index; PIE.

1 Introduction

In statistical data analysis, it is very unlikely that only one model needs to be considered. When multiple plausible models are present, the traditional approach is to take a reasonable model selection process (formally or informally) to find a single, hopefully the "best" model, from which one makes the final statistical estimation and/or prediction.

A large amount of work has been done on the topic of model selection. Various model selection methods are available based on different guiding principles and/or specific theoretical/empirical considerations, such as AIC (Akaike (1973)), BIC (Schwarz (1978)), cross-validation (Allen (1974), Stone (1974)) and MDL (e.g., Rissanen (1984) and Barron, Rissanen and Yu (1998)). Appropriate convergence properties (such as consistency of the selected model and convergence of the corresponding estimator) have been well established in a variety of settings.

*Zheng Yuan is a graduate student in Department of Biostatistics, University of Michigan (E-mail: yuanz@umich.edu). Yuhong Yang is associate professor, School of Statistics, University of Minnesota, 224 Church Street, Minneapolis 55455 (E-mail: yyang@stat.umn.edu). This work was supported by the US National Science Foundation CAREER Award Grant DMS0094323, and was mostly finished when both authors were with Department of Statistics at Iowa State University. The authors thank the editor and an associate editor for many helpful suggestions and comments.

Despite the theoretical and methodological advancement on model selection, potential problems have long been recognized. In recent years, serious concerns about the general approach of model selection have been strongly voiced. The main concern is that the uncertainty in model selection is basically ignored once a final model is found (see, e.g., Draper (1995) and Chatfield (1995)). A possible consequence is that the inference based on the final model may give an overly optimistic or misleading answer due to the under-estimation of the uncertainty associated with the whole estimation procedure (in which model selection is a non-negligible part).

Methods to address the issue have been proposed from different perspectives, including Bayesian model averaging (BMA) (or empirical Bayesian model averaging, EBMA) and weighting based on bootstrap or perturbation (see, e.g., Breiman (1996a, 1996b) and Buckland, Burnham and Augustin (1997) and references therein). The commonality is that these methods avoid selecting one model by averaging or combining the candidate models, and they have been demonstrated empirically to perform better than model selection in certain aspects for some examples.

From a Bayesian point of view, BMA is a natural approach. When a number of candidate predictors are available, a complete Bayesian solution presents challenges in terms of computation. Madigan and York (1995) proposed a Markov Chain Monte Carlo approach (called MC^3) to directly approximate the exact solution. The readers are referred to a review paper by Hoeting, Madigan, Raftery and Volinsky (1999) for details and many references on the active research in BMA.

Apart from computational challenges, there are other difficult issues for BMA. With a large number of candidate models (e.g., in the context of all subset regression with a number of predictors but a relatively small or moderate number of observations), the assignment of priors on the candidate models can be highly sensitive (cf. Fernández, Ley and Steel (2001)). For such cases, the role and meaning of the calculated posterior probabilities of the models are not quite clear, and examining frequentist properties of the BMA methods is viewed by many to be valuable (cf. George (2000)). In addition, when all the models are mis-specified (which some researchers would argue to be almost always the case), the small sample effect of weighting by the posterior probability calculated from the incorrectly specified models (and priors) seems even less clear. Thus having a set of weights (posterior probabilities) on the models does not necessarily mean that the uncertainty in model selection is properly or sufficiently taken care of. For example, when the posteriors are used to construct prediction intervals, for a crime data set, Raftery, Madigan and Hoeting (1997) applied EBMA to obtain a better predictive performance than model selection methods, but the predictive coverage was only about 80%, which was substantially lower than the intended 90% level.

More recently, Yang (2001, 2003) proposed a model combining method ARM (adaptive regression by mixing). It was shown that under some conditions, the resulting estimator performs optimally in

rate of convergence under a global L_2 loss without knowing which of the original procedures works the best. Simulations were conducted to compare ARM with model selection and BMA in Yang (2003) for a few settings. It was shown that ARM performed better than AIC, BIC and a BMA method based on BIC approximation. Regarding the convergence of BMA, unlike ARM, Yang (2004b) showed that even in simple linear regression, BMA estimators of the regression function cannot be minimax-rate optimal. Other recent non-Bayesian model averaging methods include FMA by Hjort and Claeskens (2003), where limiting distributions and risk properties of the combined estimators were established under a local asymptotic framework; and also include a very interesting method based on unbiased risk estimation by Leung and Barron (2004), which yields sharp risk bounds without requiring data splitting.

The reported empirical success of model combining shows its great potential and also calls for more work on the topic. Some of the issues that are important and of interest to us are given below.

1. Is model combining always better than model selection? If not, when is combining better than selection? Some researchers seem to suggest that since model averaging techniques (e.g., BMA) take into account the uncertainty in model selection, therefore model averaging is superior. We feel that the issue is much more complicated than that and needs to be systematically studied.

2. How should we measure the uncertainty in model selection objectively? What do we mean exactly by *uncertainty in model selection*? Obtaining a set of weights (or probabilities) on the models does not automatically mean that the uncertainty in finding the right model is correctly captured. Also, it seems desirable to have uncertainty measures that are specific to the objectives of the analysis. For example, if our interest is prediction, if several models have pretty much the same prediction performance, then which of them is selected is not the real concern at all.

3. Does there exist a clear relationship between a proper uncertainty measure of model selection and the relative performance of model combining over model selection? If so, the uncertainty measure can be used to guide us for deciding which way to go: combining or selection.

4. Should a screening step be used to eliminate poor models before combining? What are the effects of such a screening? A theoretical understanding would be helpful.

In our view, even though the concept of uncertainty in model selection is now widely recognized, the current practice of simply relying on a final model from an automated model selection by many statistics users will be improved only when the basic issues on measuring model selection uncertainty and comparison between model selection and model combining are well understood.

There is no doubt that identifying the important variables is useful for many statistical applications. It seems clear that the question of which variables are important cannot be separated from the task of

building a sensible estimator/predictor of a quantity of interest, whether a parametric or a nonparametric approach is taken. As many researchers have already pointed out, we agree, for subset selection with a large number of predictors and a small or moderate sample size, the goal of finding a single set of "important variables" is not a feasible task in general.

In this paper, we will focus on estimating the regression function with subset models and will not address issues such as confidence bands, which model is more likely to be the true one, or which set of variables is most important.

The objective of this paper is three-fold. First, we propose an instability measure to quantify the instability of model selection based on data perturbation. The idea is to have a sensible measure that can help one decide if model selection is having difficulty (hence alternatives to model selection should be considered). It is demonstrated that combining models indeed can substantially reduce instability due to model selection. A rule of thumb for using this measure is also given at the end of simulation results. Second, a combining method, ARMS, is proposed. It has a model screening step, which narrows down the list of candidate models and thus not only saves computation time but also removes very poor models that would hurt the combined estimator. A theoretical result on AMRS is presented. Third, relative performance among model combining methods and model selection methods and its relation to model selection instability are investigated in various specific cases and random settings in simulations and also some real data examples.

There are two distinct goals in combining models/procedures, one is performing as well as the best candidate model/procedure and the other is improving on the best candidate model/procedure. Our focus in this work is the former. For some discussion and references on the latter, see Yang (2003, 2004a) and references therein. Clarke (2003) gave numerical comparisons of methods with the different goals when none of the candidate models is correct.

The paper is organized as follows. We set up the problem in Section 2. In Section 3, we propose a model selection instability index (PIE) and study which factors affect it. In Section 4, we propose the ARMS algorithm with model screening before combining models and give a theoretical result. In Section 5, we compare ARMS with an empirical Bayesian model averaging (EBMA) approach and model selection methods via simulation, and propose a rule of thumb for using PIE to decide whether to combine or to select. In Section 6, we compare ARMS with EBMA, AIC and BIC in real data examples. Conclusions are in Section 7. Section 8 contains the proof of the theorem in Section 4 and a brief description of the real data sets used in the empirical study in this work.

2 Problem setup

Consider the regression problem with n observations:

$$Y_i = f(\mathbf{X}_i) + \epsilon_i, \quad i = 1, ..., n,$$

where Y is the response variable, $\mathbf{X} = (X_1, ..., X_d)$ is the explanatory variable of dimension d, $f(\cdot)$ is the true regression function, and ϵ is the random error, assumed to be normally distributed with mean 0 and variance σ^2 in this work. Throughout the paper, the observations $\mathbf{X}_i$, $1 \leq i \leq n$ are assumed to be independent of each other. For the theoretical result, the explanatory variables are further assumed to be iid and independent of the errors. For estimating f, suppose that K linear models are considered as candidates for fitting the data. The model j is

$$Y_i = f_j(\mathbf{X}_i; \theta_j) + \epsilon_i, \quad i = 1, ..., n,$$

where $f_j(\mathbf{x}; \theta_j) = \sum_{l=1}^{m_j} \theta_{j,l} \varphi_{j,l}(\mathbf{x})$ with m_j being the number of linear terms in model j, $\varphi_{j,l}(\mathbf{x})$ ($1 \leq l \leq m_j$) being the basis functions and $\theta_{j,l}$ being the linear coefficients. Let Γ denote the set of all the candidate models being considered.

Note that the model combining method to be proposed in Section 4 actually works more generally for both linear and nonlinear models. Indeed, Theorem 1 in Section 4 (the main theoretical result in this paper) is not limited to the linear model case. However, for a focused investigation, together with the fact that subset selection is a frequently encountered problem in data analysis, the rest of the paper (especially the empirical studies) will mainly deal with the subset models from d variables (i.e., all the linear models with the terms in a subset of $\{X_1, X_2, ..., X_d\}$). Clearly, there are 2^d such subset models (including the trivial model $Y_i = \text{intercept} + \epsilon_i$).

In this paper, for the theoretical work and simulations, the comparison of different estimators is under the squared L_2 loss. The squared L_2 risk of an estimator $\hat{f}$ is

$$R(f, \hat{f}) = E \parallel f - \hat{f} \parallel^2 = E \int \left(f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right)^2 P_{\mathbf{X}}(d\mathbf{x}),$$

where P_X denotes the distribution of X and the second expectation is taken with respect to the data $Z^n = (X_i, Y_i)_{i=1}^n$ under the true model.

For the empirical comparison of model selection and model combining using real data sets, we consider a random data splitting and use a predictive mean squared error (PMSE) as an objective measure of an estimation procedure (see Section 6 for details).

3 Measuring model selection instability

Clearly, regression analysis can have different goals. One possible interest is the estimation of the regression function, as is the focus of this paper. Another direction is the identification of important

explanatory variables. It seems clear that a single selected model, if reliable, is more informative than a combined estimate from different models. If there is little instability/uncertainty in a well-grounded model selection process, the selected model is most likely trustworthy; on the other hand, if there is much instability/uncertainty in model selection, the goal of identifying the "correct model" may not be realistic and insisting on simple interpretability does not seem to be appropriate. Following this consideration, an appropriate measure or index of model selection instability/uncertainty that can help one decide which scenario the data set is in is very valuable for a good statistical analysis.

It is natural to use bootstrap resampling methods to get a sense of instability in model selection (see, e.g., Breiman (1996a), Buckland *et al.* (1997) and references therein). In this paper, following Breiman (1996b), we consider an alternative based on data perturbation.

3.1 Perturbation instability in estimation (PIE)

The idea of perturbation instability is very simple: if a statistical model selection procedure is stable, a minor perturbation of the data should not change the outcome drastically. After all, there are random errors in the observed responses. Breiman (1996b) used perturbations to compare instabilities of regression procedures and also get different versions of estimators to be aggregated into a final estimator/predictor for better performance. Our use of perturbation here focuses on measuring the instability of a regression procedure quantitatively.

Consider a model selection criterion in the linear regression framework. We generate a new set of perturbation errors W_i iid from $N(0, \rho^2\hat{\sigma}^2)$, where ρ is between 0 and 1, and $\hat{\sigma}^2$ is an estimate of σ^2 based on the selected model. Note that ρ is the perturbation size, indicating the noise level of the added errors relative to the (estimated) original one. Now consider $\tilde{Y}_i = Y_i + W_i$ for $1 \le i \le n$ and apply the model selection criterion to the perturbed data set $(\tilde{Y}_i, \mathbf{X}_i), 1 \le i \le n$. If the model selection criterion is stable for the original data, then when ρ is small, the newly selected model is most likely the same as before and the corresponding estimate of f should not change too much either. At each ρ, we generate perturbation errors $\{W_i\}_{i=1}^n$ a large number of times (say 100) independently and apply the model selection procedure for each set of perturbed data. While there are different possible directions for defining perturbation instability, we here focus on perturbation instability in estimation called PIE.

At each perturbation size ρ, compute the average deviation of the new estimates of the regression function at the observed explanatory variable values from the original estimates (which are obtained from the initially selected model) based on a large number (say M) of replications:

$$I(\rho) = \frac{1}{M}\sum_{j=1}^{M} \frac{\left(\sum_{i=1}^{n}\left(\tilde{f}_j(\mathbf{X}_i) - \hat{f}(\mathbf{X}_i)\right)^2/n\right)^{1/2}}{\hat{\sigma}}, \tag{1}$$

where $\hat{f}$ and $\hat{\sigma}$ are obtained from the original data (based on the selection model) and $\tilde{f}_j$ is obtained by

applying the selection procedure again on the j-th perturbed data set. Note that all the estimates are evaluated at the original $\mathbf{X}_i$'s. The expression is similar in some aspects to Cook's distance for assessing local influence (e.g., Cook and Weisburg (1982), Cook (1986)). From the definition, $I(\rho)$ reasonably reflects the effect of perturbation on estimating f using the model selection method. Note also that the choice of the instability evaluation in (1) is only one among many possibilities. If one is interested in estimating f at a point, for example, a corresponding modification can be made. How fast the quantity $I(\rho)$ in (1) increases in ρ is a suitable instability measure of the estimation process. We plot $I(\rho)$ versus the perturbation size ρ.

Definition 1: The perturbation instability in estimation (PIE) is defined to be the slope of the perturbation plot at $\rho = 0$, (i.e., $I'(\rho)|_{\rho=0}$).

An interpretation of PIE is that if the selected model is trustworthy, when the original noise level σ is increased to about $\sqrt{1+\rho^2}\sigma$, the regression estimate changes by $PIE \cdot \hat{\sigma} \cdot \rho$ (in an average deviation sense). Note that the standard error (at least for estimating a functional) of a parametric estimator is roughly a multiple of $\hat{\sigma}/\sqrt{n}$. Thus when PIE is large, it indicates that the model selection process has produced a change at a scale more than expected, which consequently provides evidence against the reliability of the model selection procedure.

For computing PIE, we consider equally spaced ρ values with width 0.05 between 0 and 1 and then use linear regression through origin to estimate the slope of the function $I(\rho)$ at zero.

Note that the concept of PIE is not limited to model selection. As long as a regression procedure provides an estimate of σ, we can compute PIE of the procedure in the same way as above.

In the following two subsections, we study PIE with simulation and examples.

3.2 Which factors may affect PIE?

For a focused presentation, we here choose BIC as a representative of model selection methods. By our experience, the patterns for some other model selection criteria (e.g., AIC) are more or less similar. In the simulations in the following subsections, unless stated otherwise, 1) PIE refers to the perturbation instability in estimation for BIC; 2) there are 10 independent candidate predictors that are uniformly distributed on $[-1, 1]$; 3) $\sigma^2 = 1$; 4) the default sample size is 100. The PIE value for each case is the average over 50 replications.

3.2.1 Error variance σ^2

We consider two cases at various variance levels.

- *Case 3.2.1.1.* 8 predictors in the true model:

$$Y = 0.9 + 1.5X_1 + 1.6X_2 + 1.7X_3 + 1.5X_4 + 0.4X_5 + 0.3X_6 + 0.2X_7 + 0.1X_8 + \epsilon \qquad (2)$$

- *Case 3.2.1.2.*: 5 predictors in the true model:

$$Y = 1.0 + 1.0X_1 + 1.0X_2 + 1.0X_3 + 1.0X_4 + 1.0X_5 + \epsilon \tag{3}$$

The PIE values are reported in Table 1. To give a sense of variability of PIE in different replications, the standard deviations of PIE in the 50 replications are also reported in the parentheses.

	$\sigma^2 = 0.01$	0.1	0.5	1.0	2.25	4.0
Case 3.2.1.1	0.032 (0.004)	0.117 (0.023)	0.326 (0.069)	0.499 (0.100)	0.747 (0.163)	0.865 (0.210)
Case 3.2.1.2	0.029 (0.005)	0.084 (0.014)	0.214 (0.046)	0.309 (0.071)	0.535 (0.119)	0.840 (0.155)

Table 1: PIE and Error Variance: Cases 3.2.1.1 and 3.2.1.2

The table clearly shows that PIE increases as error variance σ^2 increases, which matches the intuition that model selection instability becomes larger when error variance is higher. Note that, not surprisingly, the standard deviations (in the parentheses) also increase as σ^2 increases.

3.2.2 Model complexity

In linear regression, both the number of candidate predictors and the number of predictors in the true model seem to be related to the complexity of a model selection process. We consider several cases accordingly.

The number of candidate predictors

- *Case 3.2.2.1.* Assume there are only 3 candidate predictors (X_1, X_2 and X_3) and the true model is:

$$Y = 1.0 + 1.0X_1 + 1.0X_2 + 1.0X_3 + \epsilon. \tag{4}$$

- *Case 3.2.2.2.* The true model is the same as above, but the candidate predictors are $X_1, ..., X_{10}$.

The value of PIE is 0.192 for Case 3.2.2.1 and is 0.272 for Case 3.2.2.2. As expected, the instability index of the case with more candidate predictors is bigger than that with fewer candidate predictors.

The number of predictors in the true model

- *Case 3.2.2.3.* 8 predictors in the true model:

$$Y = 1.0 + 1.0X_1 + 1.0X_2 + 1.0X_3 + 1.0X_4 + 1.0X_5 + 1.0X_6 + 1.0X_7 + 1.0X_8 + \epsilon \tag{5}$$

- *Case 3.2.2.4.* 5 predictors given by (3).

- *Case 3.2.2.5.* 3 predictors given by (4).

The values of PIE are 0.375, 0.309, 0.272 for the three cases respectively.

3.2.3 Sample size

Consider the true model in (3) with error variance 2.25. When $n = 100$, PIE is 0.535; and when n is reduced to 30, PIE is 0.756. The result agrees with the intuition that a data set with smaller sample size tends to have a bigger instability in model selection.

3.2.4 Real data examples

Here we compute PIE for several real data examples. Data sets A and B were used in the examples in several BMA papers. Short descriptions of the data sets are in Appendix 8.2. We only mention that data set A has 15 candidate predictors and 47 observations; B has 13 candidate predictors and 251 observations; C has two predictors (the original predictor and the quadratic term) and 222 observations; D has 4 candidate predictors and 32 observations; E has 7 candidate predictors and 19 observations. The data set B$'$ contains 50 observations randomly selected from the original ones in B. The second and third rows of Table 2 give the PIE values of BIC and AIC for the data sets.

	A	B	B$'$	C	D	E
BIC	0.819	0.574	0.729	0.317	0.434	0.740
AIC	0.784	0.559	0.707	0.319	0.427	0.728
ARMS	0.518	0.409	0.476	0.331	0.381	0.493
EBMA	0.537	0.417	0.524	0.322	0.388	0.539

Table 2: PIE of AIC, BIC, ARMS, and EBMA for the Real Data Sets

From the table, data set B$'$ has a significantly larger PIE value than B, and C and D have smallest PIE values.

The perturbation instability plots for A and D (as examples) are in Figures 1 and 2, respectively.

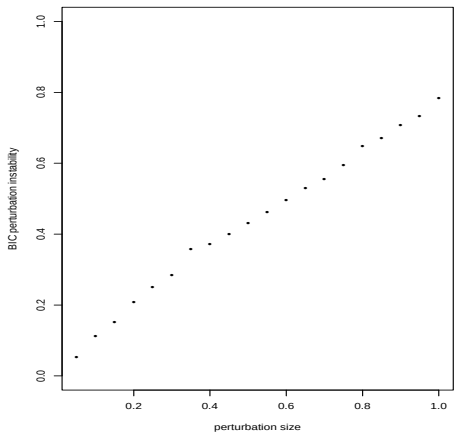

Figure 1: Perturbation Instability in Estimation for Data Set A

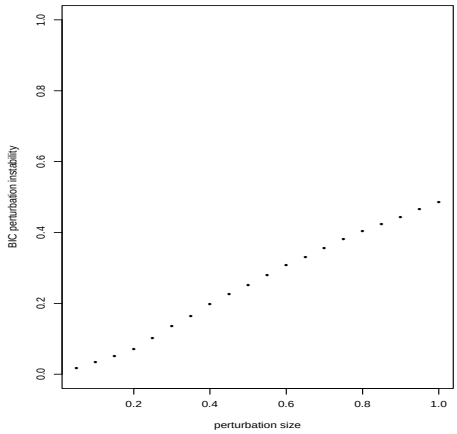

Figure 2: Perturbation Instability in Estimation for Data Set D

3.3 Combining models reduces PIE

In the previous subsection, we have seen larger instability in model selection for some cases. Here we show that combining models appropriately can reduce the instability caused by model selection. We use ARMS (to be proposed in Section 4) and EBMA (see Section 5) as model combining methods. The PIE values of ARMS and EBMA for the real data sets are given in the last two rows in Table 2. Note that for ARMS and EBMA, we estimate σ^2 by averaging those from different models using the corresponding weights (posterior probabilities for EBMA).

Table 2 shows that ARMS and EBMA both substantially reduce the instability for data sets A, B$'$ and E, where the PIE values of the model selection methods are relatively high; they significantly reduce the instability for the data set B, where PIE values of BIC and AIC are moderate; and slightly increased or decreased PIE values in data sets C and D, where the PIE values of model selection are relatively small. As will be seen in simulations and data examples later in Sections 5 and 6, model combining tends to be advantageous for cases with high PIE values.

4 Combining models by ARMS

4.1 The algorithm of ARMS with model screening

Yang (2001) proposed a method ARM (adaptive regression by mixing) to combine multiple regression models (or procedures). He examined its theoretical convergence properties and empirically demonstrated its adaptation ability in nonparametric estimation with a small number of candidate procedures. In this work, to deal with a large number of candidate parametric models, we propose an improvement of ARM with model screening. That is, we do not include all candidate models for combining. Instead,

model selection criteria AIC and BIC are used to find good candidate models for combining. This modification is based on both theoretical and practical considerations. The reduction of the number of models to be combined substantially reduces the computation cost and it can also have an advantage from a theoretical point of view (see the next subsection).

There are three main steps involved in the assignment of weights of the models for the new version of ARM. At the first step, half of the sample is used to estimate the parameters for each model j. At the second step, model selection criteria AIC and BIC are used to select a number of most promising candidate models and only these models are to be combined. At the third step, the remaining half of the response values are predicted based on the fitted models and the predictions are assessed by comparing the predicted values with the true ones. Then the models are appropriately weighted according to the performance assessment. We call this method ARMS (ARM with model Screening). For simplicity, assume that the sample size n is even. Let m be a pre-specified integer. The following is the ARMS algorithm.

- *Step 1.* Split the data into two parts $Z^{(1)} = (\mathbf{X}_i, Y_i), 1 \le i \le n/2$ and $Z^{(2)} = (\mathbf{X}_i, Y_i), n/2+1 \le i \le n$.

- *Step 2.* Estimate θ_j by $\hat{\theta}_j$ using the least squares method based on $Z^{(1)}$ for each candidate model j and compute an estimate of σ^2 (also based on $Z^{(1)}$), say $\hat{\sigma}_j^2$. Let $\hat{f}_j(\mathbf{x}) = \hat{f}_j(\mathbf{x}; \hat{\theta}_j)$.

- *Step 3.* Compute the model selection criterion values of AIC and BIC for each model j based on $Z^{(1)}$ and keep the top m models under each of the two criteria. Let Γ_s denote the set of these models (note that the size of Γ_s may be less than $2m$).

- *Step 4.* Assess the accuracies of the models using the remaining half of the data $Z^{(2)}$. For each model $j \in \Gamma_s$, predict Y_i by $\hat{f}_j(\mathbf{X}_i)$ for $n/2+1 \le i \le n$. Compute an overall measure of discrepancy:

$$D_j = \sum_{i=n/2+1}^{n} (Y_i - \hat{f}_j(\mathbf{X}_i))^2.$$

- *Step 5.* Compute the weight for model j:

$$W_j = \frac{(\hat{\sigma}_j)^{-n/2} \exp\left(-\hat{\sigma}_j^{-2} D_j/2\right)}{\sum_{k\in\Gamma_s} (\hat{\sigma}_k)^{-n/2} \exp\left(-\hat{\sigma}_k^{-2} D_k/2\right)}.$$

  Note that $\sum_{j\in\Gamma_s} W_j = 1$.

- *Step 6.* Randomly permute the order of the data $N-1$ times. Repeat the above steps and let $W_{j,r}$ denote the weight of model j computed at the r-th permutation for $0 \le r \le N-1$. Let $\hat{W}_j = \frac{1}{N}\sum_{r=0}^{N-1} W_{j,r}$

- *Step 7.* Let

$$\hat{f}_n(\mathbf{x}) = \sum_{j \in \Gamma_s} \hat{W}_j \hat{f}_j(\mathbf{x}) \tag{6}$$

be the final ARMS estimator of the true regression function f. Note that it is a convex combination of the original estimators using the models in the reduced list Γ_s.

Screening by AIC and BIC can remove some very poor models which would hurt the combined estimator. While various approaches can be considered for the choice of m (possibly data dependent), we will simply take m to be 40 in the empirical studies in this paper.

Note that after screening, the weights in Step 5 for the models in Γ_s can be interpreted as posterior probabilities of the models after observing the second part of the data with the uniform prior on the regression estimates from the first part of the data (Yang (2003, p. 787)).

4.2 A risk bound on ARMS

Regarding the ARM method of combining models/procedures, Yang (2001) gave a risk bound for the original version and an improvement was made in Yang (2003). The focus here is the theoretical consequence of the screening step. As far as we know, no risk bounds have been obtained to account for effect of model screening in the literature.

As in Yang (2001), for the theoretical result, we study a slightly different estimator from that given in (6). Let Γ_s be a reduced list of candidate models based on any consideration using the first half of the data (in this subsection, Γ_s is not necessarily obtained via AIC and BIC as in the previous subsection). For $i = n/2 + 1$, let $W_{j,i} = 1/K$ for $j \in \Gamma_s$ and for $n/2 + 1 < i \le n$, let

$$W_{j,i} = \frac{(\hat{\sigma}_j)^{-(i-n/2-1)} \exp\left(-\frac{1}{2\hat{\sigma}_j^2} \sum_{l=n/2+1}^{i-1} \left(Y_l - \hat{f}_j(\mathbf{X}_l)\right)^2\right)}{\sum_{k \in \Gamma_s} (\hat{\sigma}_k)^{-(i-n/2-1)} \exp\left(-\frac{1}{2\hat{\sigma}_k^2} \sum_{l=n/2+1}^{i-1} \left(Y_l - \hat{f}_k(\mathbf{X}_l)\right)^2\right)}.$$

Then define

$$\widetilde{W}_j = \frac{1}{n/2} \sum_{i=n/2+1}^{n} W_{j,i}$$

and let

$$\widetilde{f}(\mathbf{x}) = \sum_{j \in \Gamma_s} \widetilde{W}_j \hat{f}_j(\mathbf{x}) \tag{7}$$

be the combined estimator. The reasons for using (6) in practice instead of (7) are that the former is much simpler in computation and that we found the two formulas to be similar in performance in our experience.

We need basically the same two conditions on the models as in Yang (2003).

Condition 1: There exists a constant $\tau > 0$ such that with probability one, we have

$$\sup_{j\in\Gamma} \parallel \hat{f}_j - f \parallel_\infty \leq \sqrt{\tau}\sigma.$$

Condition 2: There exist constants $0 < \xi_1 \leq 1 \leq \xi_2 < \infty$ such that

$$\xi_1 \leq \frac{\hat{\sigma}_j^2}{\sigma^2} \leq \xi_2$$

with probability 1 for $j \in \Gamma$.

Let K and K_s denote the size of Γ and Γ_s respectively (note that K_s may be random). Let j_* denote the model in Γ that minimizes the risk $E \parallel \hat{f}_j - f \parallel^2$. Let $C(\xi_1, \xi_2) = (1/\xi_2 - 1 + \log \xi_2) \,/\, \left(\xi_1^2(1/\xi_2 - 1)^2\right)$.

Theorem 1: Assume that the errors are Gaussian and that Conditions 1 and 2 are satisfied. Then for any $j \in \Gamma$, the risk of the combined regression estimator using ARMS satisfies

$$\begin{aligned} E \parallel \widetilde{f} - f \parallel^2 \quad &\leq \quad \tau\sigma^2 P\left(j \notin \Gamma_s\right) + \\ &\quad (1 + \xi_2 + 9\tau/2)\left(\frac{2\sigma^2 E \log K_s}{n} + \frac{1}{\xi_1} E \parallel \hat{f}_j - f \parallel^2 + \frac{C(\xi_1,\xi_2)}{\sigma^2} E(\hat{\sigma}_j^2 - \sigma^2)^2\right). \quad (8) \end{aligned}$$

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

$$\begin{aligned} E \parallel \widetilde{f} - f \parallel^2 \quad &\leq \quad \tau\sigma^2 P\left(j_* \notin \Gamma_s\right) + \\ &\quad (1 + \xi_2 + 9\tau/2)\left(\frac{2\sigma^2 \log K_0}{n} + \frac{1}{\xi_1} E \parallel \hat{f}_{j_*} - f \parallel^2 + \frac{C(\xi_1,\xi_2)}{\sigma^2} E(\hat{\sigma}_{j_*}^2 - \sigma^2)^2\right). \quad (9) \end{aligned}$$

Note that when $\Gamma_s = \Gamma$ (i.e., there is no screening), the risk bound (9) stays the same as in Yang (2003). Otherwise, the new risk bound is significantly distinct: the additional term $\tau\sigma^2 P\left(j_* \notin \Gamma_s\right)$ reflects the price paid for screening; and the reduction from $\frac{2\sigma^2 \log K}{n}$ in the risk bound in Yang (2003) to $\frac{2\sigma^2 \log K_0}{n}$ in (9) (or $\frac{2\sigma^2 E \log K_s}{n}$ in (8)) suggests a potential advantage of screening in reducing the negative influence of poor models. In addition to the computational gain, this advantage, especially when K is large and K_s or K_0 is much smaller (while $P\left(j_* \notin \Gamma_s\right)$ being properly controlled), also supports the use of the screening step before combining.

The screening can be done in different ways, formally or informally (e.g., through graphical inspections) to eliminate poor models. Clearly, quantifying the effect of screening is very important. Even within the approach of using model selection criteria for screening, in addition to the method of choosing a fixed number of top models in terms of model selection criterion values as used in the algorithm in the previous subsection, one can also screen out models with obviously inferior criterion values, for which case the size of Γ_s is random. Ideally, how far to go in screening should be done in a way to balance the probability of capturing the best model (or one of the best) and the size of Γ_s.

The probability $P\left(j_* \notin \Gamma_s\right)$ can be appropriately bounded using results/techniques in the literature, e.g., Zhang (1993) and Guyon and Yao (1999) for parametric cases, Yang and Barron (1998), and Barron, Birgé, and Massart (1999) for nonparametric settings. The application of Theorem 1 to a specific screening will be given in the next subsection.

4.3 Screening out variables via a model selection criterion

Consider a subset model in the linear regression context. It is said to be an under-fitting model if at least one variable in the true model with non-zero coefficient is not included in the model.

When considering subset models, a model selection rule is said to be exponentially-inclusive if the probability of selecting any under-fitting model is exponentially small, i.e., the probability of not including any variable in the true model is upper bounded by $c_1 e^{-c_2 n^\beta}$ for some positive constants c_1, c_2, and β. The familiar model selection rules (e.g., AIC and BIC) are exponentially-inclusive. Indeed, for a model selection criterion of the form $-$log-likelihood $+ \lambda_n k$, where k is the model dimension and λ_n is the penalty constant, Guyon and Yao (1999) showed that under some mild conditions, the criterion is exponentially-inclusive as long as λ_n/n is upper bounded by a constant. Basically, for an exponentially-inclusive model selection rule, the probability of under-fitting is asymptotically negligible compared to that of over-fitting.

Let $\hat{\jmath}$ be the selected model based on a model selection rule, and let A denote the set of variables in the model $\hat{\jmath}$. Then let Γ_s be the collection of all the linear models with variables selected from A. Clearly, this screening step (from all the subset models to the reduced list Γ_s) just eliminates all the variables that are not viewed to be important by the model selection criterion.

When the model selection rule is exponentially-inclusive, we have $P(j_* \notin \Gamma_s) \le c_1 e^{-c_2 n^\beta}$ for some positive constants c_1, c_2, and β. Thus from Theorem 1, this screening method pays a small price. In contrast, a screening with a pre-determined small number of models can be overly aggressive and miss the true model with much higher probability. In practice, the aforementioned approach of screening can save computation substantially. For example, if the original number of variables is 20, and AIC selects a model with 7 terms, then for screening with AIC, there are only 2^7 many models to be combined instead of a much larger number of 2^{20}. Note that AIC tends to overfit and thus the terms not in the model selected by AIC are unlikely to be very helpful.

Since screening of models before a "formal" analysis is routinely done in practice, in our opinion, the effects of different approaches of screening need to be carefully studied for guiding real world statistical applications.

5 Simulation studies

Some limited simulations and data examples comparing ARM, BMA and model selection were reported in Yang (2003). The empirical results in this paper are different in several aspects. The simulations in Yang (2003) were done with only up to 5 predictors and no screening of models was conducted. Also the BMA method chosen there is based on BIC approximation while the BMA method based on MC^3 used in this work for comparison is regarded to be better.

	$\sigma^2 = 0.1$	0.5	1.0	2.25	4.0
ARMS	0.0135 (0.0006)	0.0615 (0.0027)	0.113 (0.0048)	0.210 (0.010)	0.392 (0.018)
BIC	0.0133 (0.0007)	0.0700 (0.0035)	0.128 (0.0056)	0.251 (0.014)	0.479 (0.024)
AIC	0.0127 (0.0006)	0.0614 (0.0030)	0.121 (0.0061)	0.254 (0.015)	0.496 (0.028)
Risk reduction	-6%	0%	7%	16%	18%
EBMA	0.0134 (0.0006)	0.0635 (0.0030)	0.119 (0.0053)	0.236 (0.012)	0.441 (0.022)
Risk reduction*	-1%	3%	5%	11%	11%

Table 3: Comparing ARMS with AIC, BIC, and EBMA: Case 5.1.1

In this section, unless stated otherwise, there are 10 candidate predictors that are independent and uniformly distributed between $[-1, 1]$. The global squared L_2 loss of a regression estimator is simulated as the average squared difference between the estimate and the true function at 1000 new independently drawn X values. For model screening, up to 80 promising candidate models are obtained using AIC and BIC (each recommends the top 40 models according to the criterion value). The sample size is 100 and the number of random permutations for ARMS is set to be 100. The values in the following tables are the simulated global squared L_2 risks based on 100 replications. The numbers in the parentheses are the corresponding standard errors. In the tables, "risk reduction" refers to the risk reduction of ARMS compared to the *best* of the model selection methods and "risk reduction*" refers to the risk reduction of ARMS relative to EBMA.

The BMA program based on MC^3 used in this work for comparison is in Splus written by Jennifer Hoeting (available at `http://www.stat.colostate.edu/~jah/software`). As mentioned earlier, MCMC is used for computing the posterior distribution in the program. In this BMA approach, conjugate priors are used for the parameters (normal for the coefficients and inverse-gamma for the variance). To determine the hyper-parameters in the prior, summary statistics of the data are used. Since the priors actually depend on the data, this is not a formal Bayes procedure in a strict sense and it seems unclear whether the the properties that hold for a formal Bayes procedure continue to hold approximately or not. From now on we call it empirical BMA (EBMA).

5.1 Comparing ARMS with AIC, BIC and EBMA

Case 5.1.1 (with small coefficients) The true model is:

$$Y = 0.9 + 1.5X_1 + 1.6X_2 + 1.7X_3 + 1.5X_4 + 0.4X_5 + 0.3X_6 + 0.2X_7 + 0.1X_8 + \epsilon$$

This model includes 8 predictors with four small and four large coefficients. The four predictors with small coefficients are difficult to identify by model selection methods when σ^2 is not small. The results

in Table 3 show that ARMS is superior in terms of L_2 risk over both model selection and EBMA when error variance σ^2 is bigger than 0.5.

Case 5.1.2 (all large coefficients) The true model is:

$$Y = 1.00 + 1.00X_1 + 1.00X_2 + 1.00X_3 + 1.00X_4 + 1.00X_5 + \epsilon$$

	$\sigma^2 = 0.1$	0.5	1.0	2.25	4.0
ARMS	0.0098 (0.0005)	0.0434 (0.0026)	0.0924 (0.0044)	0.200 (0.013)	0.358 (0.026)
BIC	0.0090 (0.0006)	0.0413 (0.0031)	0.0888 (0.0048)	0.215 (0.019)	0.483 (0.038)
AIC	0.0105 (0.0006)	0.0552 (0.0034)	0.1009 (0.0058)	0.210 (0.015)	0.390 (0.031)
Risk reduction	−9%	−5%	−4%	4%	8%
EBMA	0.0096 (0.0005)	0.0417 (0.0025)	0.0930 (0.0042)	0.202 (0.013)	0.369 (0.025)
Risk reduction*	−2%	−4%	1%	1%	3%

Table 4: Comparing ARMS with AIC, BIC, and EBMA: Case 5.1.2

Since this model does not have small coefficients, we expect that model selection has less difficulty in identifying the predictors and performs well as long as the error variance is not large. Indeed, as shown in Table 4, model selection does a better job than ARMS when $\sigma^2 \leq 1$. In addition, the result of ARMS is not much different from EBMA. EBMA and ARMS both have no advantage in this case when σ^2 is not large.

Case 5.1.3 (large model) The true model is:

$$Y = 1.0 + 1.8X_1 + 1.9X_2 + 2.0X_3 + 1.2X_4 + 1.5X_5 + 0.9X_6 + 0.8X_7 + 0.4X_8 + 0.3X_9 + 0.1X_{10} + \epsilon$$

	$\sigma^2 = 0.1$	0.5	1.0	2.25	4.0
ARMS	0.0136 (0.0006)	0.0713 (0.0035)	0.140 (0.0064)	0.269 (0.011)	0.509 (0.024)
BIC	0.0142 (0.0006)	0.0788 (0.0040)	0.161 (0.0074)	0.322 (0.016)	0.660 (0.031)
AIC	0.0134 (0.0006)	0.0701 (0.0035)	0.136 (0.0066)	0.281 (0.012)	0.540 (0.025)
Risk reduction	−1%	−2%	−3%	4%	6%
EBMA	0.0135 (0.0006)	0.0756 (0.0036)	0.159 (0.0067)	0.330 (0.014)	0.595 (0.027)
Risk reduction*	−1%	6%	12%	18%	14%

Table 5: Comparing ARMS with AIC, BIC, and EBMA: Case 5.1.3

For this case, AIC has an advantage over BIC since the true model is the full model and AIC has no chance to overfit. ARMS is not expected to be advantageous when error variance is not large. Indeed,

from Table 5, ARMS performs worse than AIC when σ^2 is less than 2. EBMA performs not very well compared to ARMS for this large model. The maximum risk ratio of EBMA over ARMS reaches 1.18 when $\sigma^2 = 2.25$.

Case 5.1.4 (small model) The true model is:

$$Y = 1.0 + 0.8X_1 + 0.9X_2 + \epsilon$$

	$\sigma^2 = 0.1$	0.5	1.0	2.25	4.0	6.25
ARMS	0.0055 (0.0004)	0.0282 (0.0010)	0.0585 (0.0044)	0.141 (0.011)	0.262 (0.015)	0.349 (0.020)
BIC	0.0050 (0.0005)	0.0268 (0.0012)	0.0573 (0.0057)	0.156 (0.012)	0.322 (0.022)	0.411 (0.027)
AIC	0.0071 (0.0006)	0.0354 (0.0016)	0.0641 (0.0061)	0.169 (0.015)	0.347 (0.020)	0.432 (0.027)
Risk reduction	−10%	−5%	−2%	10%	19%	15%
EBMA	0.0052 (0.0004)	0.0267 (0.0009)	0.0556 (0.0042)	0.135 (0.011)	0.258 (0.014)	0.350 (0.020)
Risk reduction*	−6%	-6%	−5%	−4%	−2%	0%

Table 6: Comparing ARMS with AIC, BIC, and EBMA: Case 5.1.4

This model only includes two predictors with no small coefficients. Hence, when the error variance is not large, BIC has no difficulty to select the correct model. But from Table 6, ARMS has a smaller risk than BIC and AIC when σ^2 is over 1 as the model selection instability increases. Not surprisingly, a very simple true model with a high noise level still has large instability in model selection, which makes model combining a better choice. In this case, EBMA is similar to BIC when σ^2 is small and is slightly better than ARMS when σ^2 is larger. When σ^2 reaches 6.25, however, ARMS has caught up with EBMA (in fact, for a higher σ^2, ARMS becomes better).

5.2 Random model case

The above cases were chosen to represent different scenarios. We next consider a random setting with $n = 100$ and $\sigma = 1.5$. The process is done as follows.

Case 5.2.1.

- *Step 1.* Generate the number of predictors in the true model uniformly between 1 and 10;
- *Step 2.* The coefficient of each predictor is independently generated from Uniform $(0, 2)$;
- *Step 3.* Then generate data from the model;
- *Step 4.* Compute the simulated global squared L_2 risks for the estimation procedures;
- *Step 5.* Repeat the whole process 50 times and obtain the average risk over the 50 models.

The results are summarized in Table 7, with the percentages of risk reduction of ARMS over the other procedures given. The box plot of the risks of the different methods is given in Figure 3.

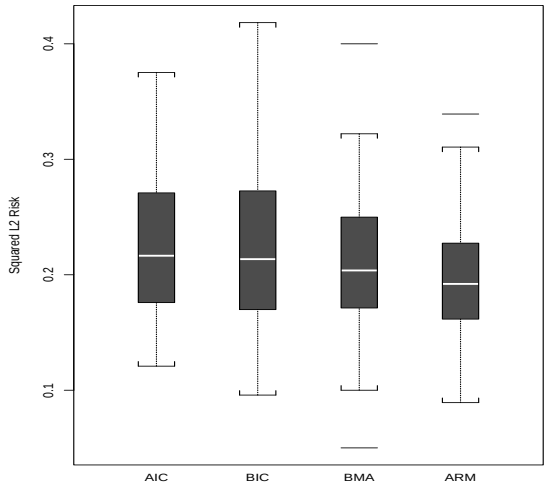

Figure 3: Random Model with Uniform Weight on the Number of Terms

	AIC	BIC	EBMA	ARMS
Average risk	0.2252 (0.0094)	0.2230 (0.098)	0.2111 (0.0082)	0.1995 (0.0075)
Risk reduction	11%	11%	5.5%	—

Table 7: Comparing ARMS with AIC, BIC, and EBMA: Case 5.2.1

To gain more insight on the relative performance of the methods, we consider two modified scenarios. For convenience, models with 1 to 5 predictors are called small models and models with 6 to 10 predictors are called large models. For both of the two cases below, each size in $\{1, .., 5\}$ has equal probability to be selected and the same is true for $\{6, ..., 10\}$.

- *Case 5.2.2.* Randomly generate models with 3/4 weight on the small models and 1/4 weight on the large models.
- *Case 5.2.3.* Randomly generate models with 1/4 weight for small models and 3/4 weight for large models.

The risks for the two cases are given in Tables 8 and 9. The box-plots for the two cases are given in Figure 4 and Figure 5.

To summarize the simulation in this subsection, both EBMA and ARMS improve over model selection

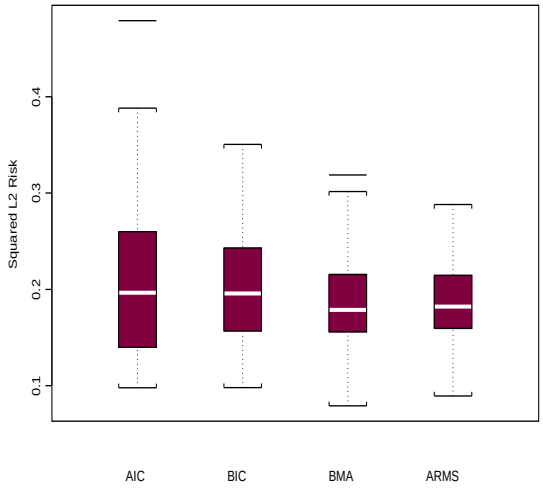

Figure 4: 3/4 Weight for Small models and 1/4 Weight for Large Models

	AIC	BIC	EBMA	ARMS
Average risk	0.2122 (0.0111)	0.2041 (0.0095)	0.1886 (0.0071)	0.1856 (0.0068)
Risk reduction	13%	9%	1.5%	—

Table 8: Comparing ARMS with AIC, BIC, and EBMA: Case 5.2.2

when the true models tend to be (relatively) small and they perform similarly. When the models tend to be larger, ARMS is superior to AIC, BIC and EBMA.

5.3 Compare ARMS with Cross-Validations

Note that ARMS is related to cross-validation in terms of data splitting and cross evaluation (though one then combines the models and the other selects one). It is natural then to compare their performance. We consider three different cross-validation methods: CV_1 (leave one out), CV_{half} (leave half out), and CV_k (leave $k = n - n^{3/4}$ out) considered by Shao (1993). With the sample size of 100, they become CV_1, CV_{50} and CV_{68} respectively.

Case 5.3.1. The true model is given in (2).

From Table 10, CV_1 is very similar to AIC and CV_{68} is very similar to BIC, as suggested by theories on models selection (see, e.g., Shao (1993)). CV_{50} is very similar to CV_{68} when error variance σ^2 is bigger than 0.5. All the three cross-validation methods perform worse than ARMS when σ^2 is not very small. Risk reductions of ARMS over the best of the CV methods are also given in the table.

Case 5.3.2. Random model case.

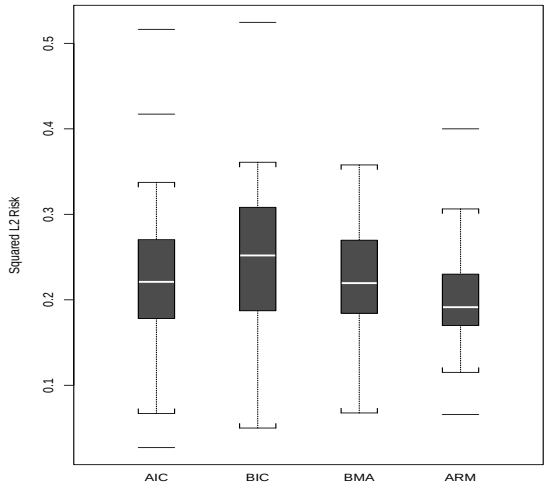

Figure 5: 1/4 Weight for Small Models and 3/4 Weight for Large Models

	AIC	BIC	EBMA	ARMS
Average risk	0.2306 (0.0092)	0.2489 (0.0108)	0.2289 (0.0089)	0.2101 (0.0081)
Risk reduction	8.5%	15.5%	8%	—

Table 9: Comparing ARMS with AIC, BIC, and EBMA: Case 5.2.3

It is conducted in a way similar to the initial random setting in Section 5.2, except that $\sigma^2 = 2$. The box plot for comparing ARMS with the cross-validation methods is in Figure 6.

Based on Table 11, risk reduction of ARMS over the cross validation methods CV_{68}, CV_{50}, and CV_1 are 14%, 10%, and 15% respectively.

5.4 Relationship between PIE and the relative performance of combining versus selection

In Section 3, we considered instability measures of model selection. Comparing the risks and PIE for the examples given there, we noticed that ARMS performed better when PIE of AIC or BIC was bigger than 0.5 and model selection tended to perform better when PIE was less than 0.4. Based on this and some additional simulations, a rule of thumb for the use of PIE for regression estimation is given below.

> *If PIE values of model selection methods are bigger than 0.5, model combining methods should be considered; if they are less than 0.4, model selection methods are OK; if PIE values are between 0.40 and 0.50, we should at least be careful about using model selection methods.*

	$\sigma^2 = 0.1$	0.5	1.0	2.25	4.0
CV_1	0.0128 (0.0006)	0.0614 (0.0031)	0.122 (0.0056)	0.256 (0.012)	0.495 (0.024)
AIC	0.0127 (0.0006)	0.0614 (0.0030)	0.121 (0.0059)	0.254 (0.015)	0.496 (0.028)
CV_{50}	0.0126 (0.0006)	0.0656 (0.0032)	0.126 (0.0055)	0.251 (0.013)	0.485 (0.024)
CV_{68}	0.0133 (0.0007)	0.0751 (0.0036)	0.127 (0.0057)	0.252 (0.015)	0.472 (0.024)
BIC	0.0133 (0.0007)	0.0720 (0.0035)	0.128 (0.0056)	0.251 (0.014)	0.479 (0.024)
ARMS	0.0135 (0.0006)	0.0615 (0.0027)	0.112 (0.0048)	0.210 (0.010)	0.392 (0.018)
Risk reduction	-6%	0%	7%	16%	18%

Table 10: Comparing ARMS with CV: Case 5.3.1

	AIC	BIC	CV_{68}	CV_{50}	CV_1	ARMS
Average risk	0.2252 (0.0094)	0.2230 (0.0098)	0.2314 (0.0134)	0.2219 (0.0087)	0.2342 (0.0089)	0.1995 (0.0077)

Table 11: Comparing ARMS with CV: Case 5.3.2

6 Data examples

In this section, we compare ARMS with EBMA, AIC and BIC using some real data sets. The performance comparison is done as follows.

First, randomly permute the order of the observations and then split the data into two parts, with the first part (n_1 observations) used for estimation, and the second part ($n - n_1$ observations) as the validation set for assessment. Second, compute the predictive mean squared error (PMSE)

$$PMSE = \frac{\sum_{i=n_1+1}^{n}(Y_i - \hat{Y}_i)^2}{n - n_1},$$

where $\hat{Y}_i$ is based on the first part of the data using the method being evaluated. Third, repeat above steps 100 times and obtain the average PMSE over the permutations.

The results are given in Table 12. The number of explanatory variables (Column 2) and split proportion $n_1 : (n - n_1)$ (Column 3) are also included in the table. For columns 5-8, the number in the parentheses is the standard error and the percentage is the risk reduction by ARMS over the method.

Below are some highlights of the results.

1. For Crime data A, the EBMA method based on MC^3 approach is significantly better than model selection AIC and BIC, but ARMS further improves the prediction accuracy by about 9%.

2. Comparing the results for data B and data B$'$, we see that when the sample size is reduced, PIE is increased and the advantage of ARMS is increased.

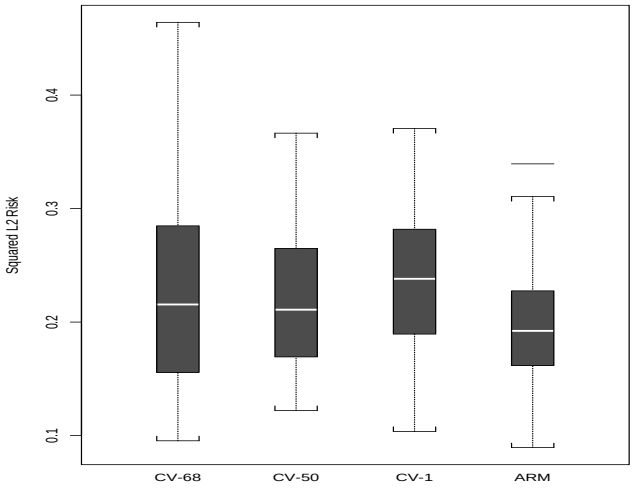

Figure 6: Comparing ARMS with Cross-Validation: Case 5.3.2

3. For Geyser data C and gas vapor data D, the risks of both ARMS and EBMA are slightly bigger than that of the model selection methods AIC and BIC.

4. Based on the table, for the six data sets, the instability measure PIE seems to work quite well as a sensible indicator of instability in model selection. For the cases with very high PIE (A, B$'$ and E), we see the advantage of ARMS over EBMA.

7 Conclusion

Both model selection and model combining have their places in statistical data analysis. However, when and how to use model combining instead of model selection is a challenging problem. We studied a model combining method with model screening and conducted a number of simulations for comparing the different approaches in representative cases and random settings.

- We proposed an index, PIE, to measure model selection instability in estimation. A rule of thumb is: If PIE is bigger than 0.5, model combining should be considered; if PIE is less than 0.4, a good model selection method is likely to work better than the model combining methods. The results from both simulations and data examples support this rule.

- We proposed a method for combining models, ARMS, with a step of model screening to remove very poor models that would hurt the combined estimator. This also saves computational cost.

- Although ARMS and EBMA both improve the estimation accuracy when PIE is high, the empirical comparisons showed that ARMS performed better than the EBMA method based on MC^3 when

	d	Split	PIE	EBMA	BIC	AIC	ARMS
A	15	37:10	0.819	0.0699 (0.0025) 9%	0.0764 (0.0029) 19%	0.0741 (0.0029) 15%	0.0637 (0.0025) —
B	13	200:51	0.574	18.65 (0.74) 2%	19.45 (0.87) 6%	19.04 (0.78) 4%	18.25 (0.77) —
B′	13	50:201	0.729	22.42 (0.86) 7%	24.71 (0.99) 16%	24.07 (0.91) 13%	20.76 (0.81) —
C	2	180:42	0.317	0.280 (0.011) -1%	0.279 (0.011) -2%	0.279 (0.011) -2%	0.284 (0.011) —
D	4	24:8	0.434	12.93 (0.37) 1%	12.55 (0.37) -2%	12.74 (0.35) -1%	12.83 (0.37) —
E	7	14:5	0.740	1.161 (0.041) 5%	1.344 (0.055) 18%	1.361 (0.049) 19%	1.10 (0.034) —

Table 12: Comparing ARMS with AIC, BIC and EBMA on Real Data Sets

the true model size is not very small and the error variance is not small. The risk reduction of ARMS over EBMA reached 18% in one case. It seems that EBMA tends to favor small models and performs not well when AIC significantly outperforms BIC. The simulation results with random models also give a consistent picture.

8 Appendix

8.1 Proof of Theorem 1

Fix a model $j^* \in \Gamma$. Let $n_1 = n_2 = n/2$. Define

$$p^{n_2} = \Pi_{i=n_1+1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - f(x_i))^2\right)$$

and

$$q^{n_2} = \sum_{j\in\Gamma_s} \frac{1}{K_s} \Pi_{i=n_1+1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} \exp\left(-\frac{1}{2\hat{\sigma}_j^2}(y_i - \hat{f}_j(x_i))^2\right).$$

Consider $\log\left(p^{n_2}/q^{n_2}\right)$. Assume first that $j^* \in \Gamma_s$. By monotonicity of the log function, we have

$$\begin{aligned}
\log\left(p^{n_2}/q^{n_2}\right) &\leq \log\left(\frac{\left(\Pi_{i=1}^{n}\left(2\pi\sigma^2\right)^{-1/2}\right)\exp\left(-\frac{1}{2}\sum_{i=n_1+1}^{n}\frac{(y_i-f(x_i))^2}{\sigma^2}\right)}{\frac{1}{K_s}\left(\Pi_{i=1}^{n}\left(2\pi\hat{\sigma}_{j^*}^2\right)^{-1/2}\right)\exp\left(-\frac{1}{2}\sum_{i=n_1+1}^{n}\frac{(y_i-\hat{f}_j(x_i))^2}{\hat{\sigma}_{j^*}^2}\right)}\right) \\
&= \log K_s + \frac{1}{2}\sum_{i=n_1+1}^{n}\left(\log\frac{\hat{\sigma}_{j^*}^2}{\sigma^2} + \frac{(y_i-\hat{f}_j(x_i))^2}{\hat{\sigma}_{j^*}^2} - \frac{(y_i-f(x_i))^2}{\sigma^2}\right). \qquad (10)
\end{aligned}$$

Taking expectation conditional on the first part of the data, as denoted by E_{n_1}, we have

$$E_{n_1}\left(\log\frac{\hat{\sigma}^2_{j^*}}{\sigma^2}+\frac{(y_i-\hat{f}_j(x_i))^2}{\hat{\sigma}^2_{j^*}}-\frac{(y_i-f(x_i))^2}{\sigma^2}\right)=\frac{\parallel\hat{f}_j-f\parallel^2}{\hat{\sigma}^2_{j^*}}+\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}-1-\log\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}. \tag{11}$$

Observe that

$$\begin{aligned}
q^{n_2} &= \sum_{j\in\Gamma_s}\frac{\frac{1}{K_s}}{\sqrt{2\pi\hat{\sigma}^2_j}}\exp\left(-\frac{1}{2\hat{\sigma}^2_j}(y_{n_1+1}-\hat{f}_j(x_{n_1+1}))^2\right)\\
&\quad\times\frac{\sum_{j\in\Gamma_s}\frac{\frac{1}{K_s}}{\sqrt{4\pi^2\hat{\sigma}^4_j}}\exp\left(-\frac{1}{2\hat{\sigma}^2_j}(y_{n_1+1}-\hat{f}_j(x_{n_1+1}))^2-\frac{1}{2\hat{\sigma}^2_j}(y_2-\hat{f}_j(x_{n_1+2}))^2\right)}{\sum_{j\in\Gamma_s}\frac{\frac{1}{K_s}}{\sqrt{2\pi\hat{\sigma}^2_j}}\exp\left(-\frac{1}{2\hat{\sigma}^2_j}(y_{n_1+1}-\hat{f}_j(x_{n_1+1}))^2\right)}\times\\
&\quad\cdots\times\frac{\sum_{j\in\Gamma_s}\frac{\frac{1}{K_s}}{\Pi_{i=1}^n\sqrt{2\pi\hat{\sigma}^2_j}}\exp\left(-\sum_{i=n_1+1}^n\frac{1}{2\hat{\sigma}^2_j}(y_i-\hat{f}_j(x_i))^2\right)}{\sum_{j\in\Gamma_s}\frac{\frac{1}{K_s}}{\Pi_{i=1}^{n-1}\sqrt{2\pi\hat{\sigma}^2_j}}\exp\left(-\sum_{i=n_1+1}^{n-1}\frac{1}{2\hat{\sigma}^2_j}(y_i-\hat{f}_j(x_i))^2\right)}.
\end{aligned}$$

Let $p_i=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-f(x_i))^2}{2\sigma^2}\right)$ and $g_i=\sum_{j\in\Gamma_s}W_{j,i}\frac{1}{\sqrt{2\pi\hat{\sigma}^2_j}}\exp\left(-\frac{(y_i-\hat{f}_j(x_i))^2}{2\hat{\sigma}^2_j}\right)$. It follows by the definition of $W_{j,i}$ that $\log\left(p^{n_2}/q^{n_2}\right)=\sum_{i=n_1+1}^n\log\left(\frac{p_i}{g_i}\right)$. Together with (10) and (11), under the i.i.d. assumption on the data, we have

$$\sum_{i=n_1+1}^n E_{n_1}\log\left(\frac{p_i}{g_i}\right)\le\log K_s+\frac{1}{2}n_2E_{n_1}\left(\frac{\parallel\hat{f}_j-f\parallel^2}{\hat{\sigma}^2_{j^*}}+\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}-1-\log\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}\right). \tag{12}$$

Conditional on the first part of the data and the explanatory variables, as denoted by $E^{'}_{n_1}$ below, we have

$$E^{'}_{n_1}\log\left(\frac{p_i}{g_i}\right)=\int p_i\log\frac{p_i}{g_i}dy_i\ge\int\left(\sqrt{p_i}-\sqrt{g_i}\right)^2dy_i.$$

Let $\overline{f}_i(x)=\sum_{j\in\Gamma_s}W_{j,i}\hat{f}_j$ for $n_1+1\le i\le n$. By Lemma 1 of Yang (2001), under Conditions 1 and 2, we have that for $n_1+1\le i\le n$,

$$E^{'}_{n_1}\log\left(\frac{p_i}{g_i}\right)\ge\frac{\left(\overline{f}_i(x_i)-f(x_i)\right)^2}{\sigma^2\left(2(1+\xi_2)+9\tau\right)}.$$

Together with (12), we have

$$\sum_{i=n_1+1}^n E_{n_1}\left(\frac{\left(\overline{f}_i(x_i)-f(x_i)\right)^2}{\sigma^2\left(2(1+\xi_2)+9\tau\right)}\right)\le\log K_s+\frac{1}{2}n_2E_{n_1}\left(\frac{\parallel\hat{f}_j-f\parallel^2}{\hat{\sigma}^2_{j^*}}+\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}-1-\log\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}\right).$$

That is,

$$\frac{1}{n_2}\sum_{i=n_1+1}^n\parallel\overline{f}_i-f\parallel^2\le\sigma^2\left(2(1+\xi_2)+9\tau\right)\left(\frac{\log K_s}{n}+\frac{1}{2}E_{n_1}\left(\frac{\parallel\hat{f}_j-f\parallel^2}{\hat{\sigma}^2_{j^*}}+\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}-1-\log\frac{\sigma^2}{\hat{\sigma}^2_{j^*}}\right)\right).$$

By convexity of the squared L_2 norm, together with that $\widetilde{f}(x)=\frac{1}{n_2}\sum_{i=n_1+1}^n\overline{f}_i$, we have

$$\parallel\widetilde{f}-f\parallel^2\le\frac{1}{n_2}\sum_{i=n_1+1}^n\parallel\overline{f}_i-f\parallel^2.$$

Note that if $x \geq x_0 > 0$, $x - 1 - \log x \leq c_{x_0}(x-1)^2$ for a constant $c_{x_0} = \frac{x_0 - 1 - \log x_0}{(x_0-1)^2}$. It follows that when $j^* \in \Gamma_s$,

$$\| \widetilde{f} - f \|^2 \leq (1 + \xi_2 + 9\tau/2) \left(\frac{2\sigma^2 \log(K_s)}{n} + \frac{1}{\xi_1} \| \hat{f}_{j^*} - f \|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} (\hat{\sigma}^2_{j^*} - \sigma^2)^2 \right),$$

where $C(\xi_1, \xi_2) = \frac{1/\xi_2 - 1 + \log \xi_2}{\xi_1^2 (1/\xi_2 - 1)^2}$. Since $\widetilde{f}$ is a convex combination of the original estimators, under Condition 1, we have $\| \widetilde{f} - f \|^2 \leq \tau\sigma^2$ when $j^* \notin \Gamma_s$. Let G_s denote the event that model j^* is in Γ_s. It follows that

$$\| \widetilde{f} - f \|^2 \leq \tau\sigma^2 I_{G_s^c} + (1 + \xi_2 + 9\tau/2) \left(\frac{2\sigma^2 \log(K_s)}{n} + \frac{1}{\xi_1} \| \hat{f}_{j^*} - f \|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} (\hat{\sigma}^2_{j^*} - \sigma^2)^2 \right) I_{G_s},$$

where $I_{\{\}}$ denoe the indicator function. The conclusion then follows. This completes the proof of Theorem 1.

8.2 A brief description of the real data sets used in the paper

1. Crime data (A). The data set contains the crime rate (response) and 15 predictors of 47 states in US and was used by Ehrlich (1973) as an example to test a theoretical argument on crime. Vandaele (1978) corrected some errors and we use the corrected data in this work. This data was considered in several BMA papers (e.g., Hoeting *et al.* (1999)). As in the original analysis, all values except the indicator variable for southern states were transformed logarithmically.

2. Fat data (B and B′). This data set gives body fat measurements for 252 men. The goal of the analysis was to predict percentage of body fat using 13 simple body measurements (see Penrose, Nelson and Fisher (1985) and Johnson (1996)). For each subject, percentage of body fat, age, weight, height and ten body circumference measurements were recorded. One subject (observation 42) was removed since the height given was obviously incorrect.

3. Geyser data (C). This data set was obtained from Simonoff (1996). The data set includes one predictor: eruption duration time. The response is eruption time interval. There are 222 observations in this data set. We chose the predictor itself and its square term as our two candidate predictors.

4. Gas vapor data (D). This data set contains 4 predictors and 32 observations (see, e.g., Weisberg (1985, p. 138)). The response variable is the amount of vapor that is vented into the atmosphere when gasoline is pumped into the tank of a car.

5. Pull strength data (E). There are 19 observations. The data set consists of information on pull strength (response) of a wire bond, die height, post height, loop height, wire length, bond width on the die and bond width on the post. A product term of the second and the fourth predictors is added as the seventh candidate predictor (see, e.g., Myers, Montgomery and Vining (2002, p. 53)).

References

[1] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Info. Theory*, 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest.

[2] Allen, D. M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16, 125-127.

[3] Barron, Birgé and Massart (1999) Risk Bounds for Model Selection Via Penalization, *Probability Theory and Related Fields*, 113, 301-413.

[4] Barron, A.R., Rissanen, J. and Yu, B. (1998) The minimum description length principle in coding and modeling, *IEEE: Information Theory*, 44, 2743-2760.

[5] Breiman, L. (1996a) Bagging predictors, *Machine Learning*, 24, 123-140.

[6] Breiman, L. (1996b) Heuristics of instability and stabilization in model selection, *Annals of Statistics*, 24, 2350-2383.

[7] Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997) Model selection: an integral part of inference, *Biometrics*, 53, 603-618.

[8] Chatfield, C. (1995) Model uncertainty, data mining and statistical inference (with discussion), *Journal of the Royal Statistical Society*, Series A, 158, 419-466.

[9] Clarke, B. (2003) Comparing Bayes model averaging and stacking when model approximation error cannot be ignored, *Journal of Machine Learning Research*, 4, 683-712.

[10] Cook, R.D. (1986) Assessment of local influence (with discussion), *Journal of the Royal Statistical Society*, Series B, 48, 133-155.

[11] Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. Chapman & Hall, New York.

[12] Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion), *Journal of the Royal Statistical Society*, Series B, 57, 45-70.

[13] Ehrlich, I. (1973) Participation in illegitimate activities: a theoretical and empirical investigation, *Journal of Political Economy*, 81, 521-565

[14] Fernández, C., Ley, E. and Steel, M. F. J. (2001) Benchmark priors for Bayesian Model Averaging, *Journal of Econometrics*, 100, 381-427.

[15] George, E. (2000) The Variable selection problem, *Journal of the American Statistical Association*, 95, 1304-1308.

[16] Guyon, X. and Yao, J. (1999) On the underfitting and overfitting sets of models chosen by order selection criteria, *Journal of Multivariate Analysis*, 70, 221-249.

[17] Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1999) Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, 14, 382-417.

[18] Hjort, N.L. and Claeskens, G. (2003) Frequentist model average estimators (with discussion), *Journal of the American Statistical Association*, 98, 879-899.

[19] Johnson, R.W. (1996) Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4.

[20] Leung, G. and Barron, A.R. (2004) Information theory and mixing least-squares regressions, manuscript.

[21] Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data, *International Statistical Review*, 63, 215-232.

[22] Myers, R.H. , Montgomery, D.C. , and Vining, G.G. (2002) *Generalized Linear Models: with Applications in Engineering and the Sciences*, John Wiley & Sons, New York.

[23] Penrose, K., Nelson, A. and Fisher, A. (1985) Generalized body composition prediction equation for men using simple measurement techniques (abstract), *Medicine and Science in Sports and Exercises*, 19, 189.

[24] Raftery, A., Madigan, D. and Hoeting, J. (1997) Bayesian model averaging for linear regression models, *Journal of American Statistical Association*, 92, 179-191.

[25] Rissanen, J. (1984) Universal coding, information, prediction, and estimation, *IEEE Transactions on Information Theory*, 30, 629-636.

[26] Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statistics*, 6, 461-464.

[27] Shao, J. (1993) Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.

[28] Simonoff, J.S. (1996) *Smoothing Methods in Statistics*. Springer-Verlag, New York.

[29] Stone, M. (1974) Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, Ser.B, 36, 111-147

[30] Vandaele, W. (1978) Participation in illegitimate activities; Ehrlich revisited, In *Deterrence and Incapacitation* (eds. A. Blumstein, J. Cohen, and D. Nagin), Washington, D.C.: National Academy of Sciences Press, 270-335.

[31] Weisberg, S. (1985) *Applied Linear Regression*, John Wiley & Sons, New York.

[32] Yang, Y. (2001) Adaptive regression by mixing. *Journal of American Statistical Association*, 96, 574-588.

[33] Yang, Y. (2003) Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, 13, 783-809.

[34] Yang, Y. (2004a) Aggregating regression procedures to improve performance, *Bernoulli*, 10, 25-47.

[35] Yang, Y. (2004b) Can the strengths of AIC and BIC be shared? -A conflict between model identification and regression estimation, manuscript.

[36] Yang, Y. and Barron, A.R. (1998) Asymptotic property of model selection criteria, *IEEE Transactions on Information Theory*, 44, 95-116.

[37] Zhang, P. (1993), On the convergence rate of model selection criteria, *Commun. Statist.-Theory Meth.*, 22, 2765-2775.