Combining linear regression models: When and how?

Zheng Yuan
University of Michigan - Ann Arbor

Yuhong Yang
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/stat_las_preprints

Part of the Statistics and Probability Commons

Recommended Citation
http://lib.dr.iastate.edu/stat_las_preprints/117

This Article is brought to you for free and open access by the Statistics at Iowa State University Digital Repository. It has been accepted for inclusion in Statistics Preprints by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Combining linear regression models: When and how?

Abstract
Model-combining (i.e., 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, first we propose an instability measure to capture the uncertainty of model selection in estimation, called perturbation instability in estimation (PIE), based on perturbation of the sample. We demonstrate that estimators from model selection can have large PIE values and that model combining substantially reduces the instability for such cases. Second, we propose a model combining method, adaptive regression by mixing with model screening (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. Third, 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 that 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.

Keywords
adaptive regression by mixing, Bayesian model averaging, instability index, model combining, model selection, model uncertainty, perturbation instability in estimation

Disciplines
Statistics and Probability

Comments
Combining Linear Regression Models: When and How?

Zheng Yuan and Yuhong Yang
Department of Biostatistics, University of Michigan
and Department of Statistics, Iowa State University

June, 2003

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 BMA (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 BMA 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; PIS.

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, based on which one makes the final statistical estimation and/or prediction.

Perhaps due to the relevance of model selection in almost all statistical applications, 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)) and cross-validation (Allen (1974), Stone (1974)). 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.

Despite the theoretical and methodological advancement on model selection, potential problems
have been aware to statisticians. 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) 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 \( \text{MC}^3 \)) 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 BMA 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 showed that ARM performed better than AIC, BIC and a BMA method based on BIC approximation.

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 carefully 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 will be improved only when the basic issues on measuring model selection uncertainty and comparisons between model selection and model combining are well understood.

There is no doubt that identifying the important variables is an important issue for many statistical applications. It seems clear that the question 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 approaching 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 instability measures to quantify the instability of model selection in two different directions 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. 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 directions in combining models/procedures, one targeting at performing as well as the best candidate model/procedure and the other targeting at improving the best candidate model/procedure. Our focus in this work is the former. For some discussion and references on the latter, see Yang (2003) and references therein.

The paper is organized as follows. We set up the problem in Section 2. In Section 3, we propose model selection instability indices (PIE and PIS) and study which factors affect them. 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 BMA and model selection methods in simulation, and propose and study a rule of thumb for using PIE to help one decide whether to combine or to select. In Section 6, we compare ARMS with BMA, 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:

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $Y$ is the response variable, $X = (X_1, \ldots, X_d)$ is the explanatory variable of dimension $d$, $f(\cdot)$ is the true regression function, and $\epsilon$ is the random error, assumed to be normally distributed with mean 0 and variance $\sigma^2$ in this work. Throughout the paper, the observations $X_i, 1 \leq i \leq n$ 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

$$Y_i = f_j(X_i; \theta_j) + \epsilon_i, \quad i = 1, \ldots, n,$$
where \( f_j(x; \theta_j) = \sum_{l=1}^{m_j} \theta_{j,l} \varphi_{j,l}(x) \) with \( m_j \) being the number of linear terms in model \( j \), \( \varphi_{j,l}(x) \) (1 \( \leq l \leq m_j \)) being the basis functions and \( \theta_{j,l} \) being the linear coefficients. Let \( \Gamma \) 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 with \( d \) variables (i.e., all the linear models with the terms from a subset of \( \{X_1, X_2, \ldots, X_d\} \)). 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

\[
R(f, \hat{f}) = E \left( f(X) - \hat{f}(X) \right)^2 = E \int \left( f(x) - \hat{f}(x) \right)^2 P_X(dx),
\]

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 an average prediction 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 errors in the observed responses. Breiman (1996b) used perturbations to compare instability 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 $\rho$ is between 0 and 1, and $\hat{\sigma}^2$ is an estimate of $\sigma^2$ based the selected model. Note that $\rho$ is the perturbation size, indicating the noise level of the added errors relative to the (estimated) original one. Now consider $\hat{Y}_i = Y_i + W_i$ for $1 \leq i \leq n$ and apply the model selection criterion to the perturbed data set $(\hat{Y}_i, \mathbf{X}_i), 1 \leq i \leq n$. If the model selection criterion is stable for the original data, then when $\rho$ 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 $\rho$, 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 $\rho$, 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:

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

where $\hat{f}$ and $\hat{\sigma}$ are obtained from the original data (based on the selection model) and $\hat{f}_j$ 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. 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 $\rho$ is a suitable instability measure of the estimation process. We plot $I(\rho)$ versus the perturbation size $\rho$.

**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 $\sigma$ is increased to about $\sqrt{1 + \rho^2 \hat{\sigma}}$, 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 $\sigma/\sqrt{n}$. 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 $\rho$ 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 $\sigma$, 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 $\sigma^2$

We consider two cases at various variance levels.

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

  $$Y = 0.9 + 1.5X_1 + 1.6X_2 + 1.7X_3 + 1.5X_4 + 0.4X_5 + 0.3X_6 + 0.2X_7 + 0.1X_8 + \epsilon$$  \hspace{1cm} (2)

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

  $$Y = 1.0 + 1.0X_1 + 1.0X_2 + 1.0X_3 + 1.0X_4 + 1.0X_5 + \epsilon$$  \hspace{1cm} (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.

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>0.01</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>2.25</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 3.2.1.1</td>
<td>0.032</td>
<td>0.117</td>
<td>0.326</td>
<td>0.499</td>
<td>0.747</td>
<td>0.865</td>
</tr>
<tr>
<td>(0.004)</td>
<td>(0.023)</td>
<td>(0.069)</td>
<td>(0.100)</td>
<td>(0.163)</td>
<td>(0.210)</td>
<td></td>
</tr>
<tr>
<td>Case 3.2.1.2</td>
<td>0.020</td>
<td>0.084</td>
<td>0.214</td>
<td>0.309</td>
<td>0.535</td>
<td>0.840</td>
</tr>
<tr>
<td>(0.005)</td>
<td>(0.014)</td>
<td>(0.046)</td>
<td>(0.071)</td>
<td>(0.119)</td>
<td>(0.155)</td>
<td></td>
</tr>
</tbody>
</table>

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 $\sigma^2$ 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 $\sigma^2$ 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:

  $$Y = 1.0 + 1.0X_1 + 1.0X_2 + 1.0X_3 + \epsilon.$$  

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

  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:

  $$Y = 1.0 + 1.0X_1 + 1.0X_2 + 1.0X_3 + 1.0X_4 + 1.0X_5 + 1.0X_6 + 1.0X_7 + 1.0X_8 + \epsilon$$  

- **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 $n$ reduces 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 are 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.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>B'</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>0.819</td>
<td>0.574</td>
<td>0.729</td>
<td>0.317</td>
<td>0.434</td>
<td>0.740</td>
</tr>
<tr>
<td>AIC</td>
<td>0.784</td>
<td>0.559</td>
<td>0.707</td>
<td>0.319</td>
<td>0.427</td>
<td>0.728</td>
</tr>
<tr>
<td>ARMS</td>
<td>0.518</td>
<td>0.409</td>
<td>0.476</td>
<td>0.331</td>
<td>0.381</td>
<td>0.493</td>
</tr>
<tr>
<td>BMA</td>
<td>0.537</td>
<td>0.417</td>
<td>0.524</td>
<td>0.322</td>
<td>0.388</td>
<td>0.539</td>
</tr>
</tbody>
</table>

Table 2: PIE of AIC, BIC, ARMS, and BMA 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.

![Perturbation Instability in Estimation for Data Set A](image-url)

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 BMA (see Section 5) as model combining methods. The PIE
values of ARMS and BMA for the real data sets are given in the last two rows in Table 2. Note that for ARMS and BMA, we estimate $\sigma^2$ by averaging those from different models using the corresponding weights (posterior probabilities for BMA).

Table 2 shows that ARMS and BMA 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.

### 3.4 Perturbation instability in selection (PIS)

In the same framework considered for PIE, differently from measuring instability in estimation, one can also assess instability in terms of selection. At each perturbation size $\rho$, we record the percentage of times that the originally selected model is not chosen again with the addition of the perturbation errors in e.g., 100 replications of perturbations. We then plot the percentage versus the perturbation size $\rho$. If the percentage increases sharply in $\rho$, it indicates that the selection procedure is unstable, which in turn casts doubt on its ability of choosing the “best” model.

**Definition 2:** The perturbation instability in selection (PIS) is defined to be the slope of the above perturbation plot at $\rho = 0$.

Similar to PIE, an interpretation of PIS is that when the original noise level $\sigma$ is increased to about $\sqrt{1 + \rho^2 \sigma}$ (for a small $\rho$), the chance of selecting a different model is about $PIS \cdot \rho$. When PIS is large, the model selection procedure could quite easily choose a different model even with a small perturbation.
For example, if $PIS = 1$, with 11% increase of the noise level ($\rho = 0.5$), roughly speaking, half of the time a different model would be selected. That seems to be a clear indication that we cannot trust much the selected model.

Note that PIE and PIS are clearly related, but they can take quite different values in general. For example, consider two candidate models. If they have very similar performances in estimation, then it is quite possible that PIE is small but PIS is large.

Since selecting variables is not our main concern in this paper, we report below fewer simulations and data examples for PIS regarding which factors may affect PIS. Unless stated otherwise, the simulation settings and cases are the same as the corresponding ones used for PIE.

### 3.4.1 Error variance $\sigma^2$

We only give Case 3.2.1.2 in Table 3. The other case is similar.

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>0.01</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>2.25</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.260</td>
<td>0.313</td>
<td>0.354</td>
<td>0.466</td>
<td>0.834</td>
<td>1.181</td>
</tr>
</tbody>
</table>

Table 3: PIS and Error Variance: Case 3.2.1.2

Note that even at $\sigma^2 = 0.01$, PIS is not small, indicating the substantial instability in model selection from the selection point of view.

### 3.4.2 The number of candidate predictors

The value of PIS is 0.05 for Case 3.2.2.1 and is 0.379 for Case 3.2.2.2. Clearly, here the number of candidate variables had a much larger effect on PIS compared to PIE.

### 3.4.3 Sample size

For the same case (true model in (3)), PIS is 0.466 at $n = 100$ and increases to 0.986 at $n = 30$.

### 3.4.4 Real data examples

The PIS values of BIC for the real data sets are given in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>B'</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>1.613</td>
<td>1.058</td>
<td>1.392</td>
<td>0.085</td>
<td>0.389</td>
<td>1.310</td>
</tr>
</tbody>
</table>

Table 4: PIS of BIC for the Real Data Sets

Note that for data sets A, B, B' and E, PIS values are all above 1, strongly suggesting that we can have little confidence on the selected models. In contrast, BIC is very stable on the data set C.
3.5 A remark

We have used data perturbation to characterize the instability of a regression procedure and defined instability measures--PIE and PIS. Based on the empirical results, they seem to reflect the instability of model selection in several aspects appropriately, such as error variance, model complexity and sample size. It would be nice if we could depend on the measures to decide whether model combining is preferred to model selection or not. Our empirical studies led us to a rule of thumb to be given later in Section 5.

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, in dealing 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 $Z^{(1)} = (X_i, Y_i), 1 \leq i \leq n/2$ and $Z^{(2)} = (X_i, Y_i), n/2 + 1 \leq i \leq n$.

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

- **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 $\Gamma_s$ denote the set of these models (note that the size of $\Gamma_s$ 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 $\hat{f}_j(X_i)$ for $n/2 + 1 \leq i \leq n$. Compute an overall measure of discrepancy:

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

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

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

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 $0 \leq r \leq N - 1$. Let

$$W_j = \frac{1}{N} \sum_{r=0}^{N-1} W_{j,r}$$

- **Step 7.** Let

$$\hat{f}_n(x) = \sum_{j \in \Gamma_s} W_j \hat{f}_j(x)$$

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 $\Gamma_s$.

Screening by AIC and BIC can remove some very poor models which would hurt the combined estimator. It also has the advantage of saving computation time compared to the task of computing weights for all the candidate models. 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.

### 4.2 A theoretical result 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 in Yang (2001), for the theoretical result, we study a slightly different (more involved in computing) estimator from that given in (6). Let $\Gamma_s$ be a reduced list of candidate models based on any consideration using the first half of the data (in this section, $\Gamma_s$ 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 $n/2 + 1 < i \leq n$, let

$$W_{j,i} = \frac{(\sigma_j)^{-(i-n/2-1)} \exp \left(-\frac{1}{2\sigma_j^2} \sum_{i=n/2+1}^{i-1} \left(Y_i - \hat{f}_j(X_i)\right)^2 \right)}{\sum_{k \in \Gamma_s} (\sigma_k)^{-(i-n/2-1)} \exp \left(-\frac{1}{2\sigma_k^2} \sum_{i=n/2+1}^{i-1} \left(Y_i - \hat{f}_k(X_i)\right)^2 \right)}.$$
Then define
\[ \widehat{W}_j = \frac{1}{n/2} \sum_{i=n/2+1}^{n} W_{j,i} \]
and let
\[ \bar{f}(x) = \sum_{j=1}^{K} \widehat{W}_j f_j(x) \]  
be the combined estimator.

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
\[ \sup_{j \in \Gamma} \| \bar{f}_j - f \|_\infty \leq \sqrt{\tau \sigma}. \]

**Condition 2:** There exist constants \( 0 < \xi_1 \leq 1 \leq \xi_2 < \infty \) such that
\[ \xi_1 \leq \frac{\sigma^2}{\sigma_2^2} \leq \xi_2 \]
with probability 1 for \( j \in \Gamma \).

Let \( K \) and \( K_n \) denote the size of \( \Gamma \) and \( \Gamma_n \) respectively (note that \( K_n \) may be random). Let \( j_* \) denote the model in \( \Gamma \) that minimizes the risk \( E \| \bar{f}_j - f \|^2 \). Let \( C(\xi_1, \xi_2) = (1/\xi_2 - 1 + \log \xi_2) / (\xi_1^2 (1/\xi_2 - 1)^2) \).

**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
\[
E \| \bar{f} - f \|^2 \leq \tau \sigma^2 P(j \notin \Gamma) + (1 + \xi_2 + 9\tau/2) \left( \frac{2\sigma^2 E \log K_n}{n} + \frac{1}{\xi_1} E \| \bar{f}_j - f \|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} E(\sigma^2_j - \sigma^2)^2 \right). \tag{8}
\]

In particular, when \( K_n \) is upper bounded by a constant \( K_0 \), we have
\[
E \| \bar{f} - f \|^2 \leq \tau \sigma^2 P(j_* \notin \Gamma) + (1 + \xi_2 + 9\tau/2) \left( \frac{2\sigma^2 \log K_0}{n} + \frac{1}{\xi_1} E \| \bar{f}_{j_*} - f \|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} E(\sigma^2_{j_*} - \sigma^2)^2 \right). \tag{9}
\]

Note that when \( \Gamma_n = \Gamma \) (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 \( \tau \sigma^2 P(j_* \notin \Gamma) \) reflects the price paid for screening; and the reduction from \( \frac{2\sigma^2 \log K_n}{n} \) in the risk bound in Yang (2003) to \( \frac{2\sigma^2 \log K_0}{n} \) in (9) (or \( \frac{2\sigma^2 E \log K_n}{n} \) in (8)) suggests a potential advantage of screening in reducing the negative influence of poor models. In additional to the computational gain, this advantage, especially when \( K \) is large and \( K_n \) or \( K_0 \) is much smaller (while \( P(j_* \notin \Gamma) \) 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. Even within the approach of using model selection criteria in that
regard, 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 section, one could also screen out models with
obviously inferior criterion values, for which case the size of $\Gamma_s$ 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 $\Gamma_s$.

The probability $P(j_* \notin \Gamma_s)$ can be appropriately bounded using results/techniques in the literature,
e.g., Zhang (1993) for parametric cases, Yang and Barron (1998), and Barron, Birgé, and Massart (1999)
for nonparametric settings.

5 Simulation studies

Some limited simulations and data examples comparing ARM, BMA and models 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.

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 BMA.

The BMA program based on $MC^3$ used in this work for comparison is in Splus written by Jen-
nifer Hoeting (available at http://www.stat.colostate.edu/~jah/software). As mentioned earlier,
MCMC is used for computing the posterior distribution in the program.

5.1 Comparing ARMS with AIC, BIC and BMA

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 be identified by model selection methods when $\sigma^2$ is not small. The
results in Table 5 show that ARMS is superior in terms of $L_2$ risk over both model selection and BMA when error variance $\sigma^2$ is bigger than 0.5.

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

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

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 6, model selection does a better job than ARMS when $\sigma^2 \leq 1$. In addition, the result of ARMS is not much different from BMA. BMA and ARMS both have no advantage in this case when $\sigma^2$ is not large.

**Case 5.1.3 (large model)** The true model is:

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

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 7, ARMS performs worse than AIC when \( \sigma^2 \) is less than 2. BMA performs not very well compared to ARMS for this large model. The maximum risk ratio of BMA over ARMS reaches 1.23 when \( \sigma^2 = 2.25 \).

Case 5.1.4 (small model) The true model is:

\[
Y = 1.0 + 0.8X_1 + 0.9X_2 + \epsilon
\]

<table>
<thead>
<tr>
<th>( \sigma^2 )</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>2.25</th>
<th>4.0</th>
<th>6.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMS</td>
<td>0.0136</td>
<td>0.0142</td>
<td>0.0134</td>
<td>0.0135</td>
<td>0.0135</td>
<td>0.0135</td>
</tr>
<tr>
<td></td>
<td>(0.0006)</td>
<td>(0.0006)</td>
<td>(0.0006)</td>
<td>(0.0006)</td>
<td>(0.0006)</td>
<td>(0.0006)</td>
</tr>
<tr>
<td>BIC</td>
<td>0.0773</td>
<td>0.0788</td>
<td>0.0701</td>
<td>0.0706</td>
<td>0.0706</td>
<td>0.0706</td>
</tr>
<tr>
<td></td>
<td>(0.0040)</td>
<td>(0.0040)</td>
<td>(0.0035)</td>
<td>(0.0036)</td>
<td>(0.0036)</td>
<td>(0.0036)</td>
</tr>
<tr>
<td>AIC</td>
<td>0.140</td>
<td>0.161</td>
<td>0.136</td>
<td>0.159</td>
<td>0.159</td>
<td>0.159</td>
</tr>
<tr>
<td></td>
<td>(0.0064)</td>
<td>(0.0074)</td>
<td>(0.0066)</td>
<td>(0.0067)</td>
<td>(0.0067)</td>
<td>(0.0067)</td>
</tr>
<tr>
<td>Risk reduction</td>
<td>-1%</td>
<td>-2%</td>
<td>-3%</td>
<td>4%</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>BMA</td>
<td>0.269</td>
<td>0.322</td>
<td>0.281</td>
<td>0.330</td>
<td>0.350</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.016)</td>
<td>(0.012)</td>
<td>(0.014)</td>
<td>(0.014)</td>
<td></td>
</tr>
<tr>
<td>Risk reduction*</td>
<td>-1%</td>
<td>6%</td>
<td>12%</td>
<td>18%</td>
<td>14%</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Comparing ARMS with AIC, BIC, and BMA: 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 8, ARMS has a smaller risk than BIC and AIC when \( \sigma^2 \) 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, BMA is similar to BIC when \( \sigma^2 \) is small and is slightly better than ARMS when \( \sigma^2 \) is larger. When \( \sigma^2 \) reaches 6.25, however, ARMS has caught up with BMA (in fact, for a higher \( \sigma^2 \), ARMS becomes better).
5.2 Random model case

The above cases were chosen to represent different scenarios. In order to compare the different methods more fairly, we 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 9, 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.

![Box plot of squared risks for different methods](image)

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

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>BMA</th>
<th>ARMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.2252</td>
<td>0.2230</td>
<td>0.2111</td>
<td>0.1995</td>
</tr>
<tr>
<td></td>
<td>(0.0094)</td>
<td>(0.098)</td>
<td>(0.0082)</td>
<td>(0.0075)</td>
</tr>
<tr>
<td>Risk reduction</td>
<td>11%</td>
<td>11%</td>
<td>5.5%</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 9: Comparing ARMS with AIC, BIC, and BMA: 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 \( \{1, \ldots, 5\} \) has equal probability to be selected and the same is true for \( \{6, \ldots, 10\} \).

- **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 10 and 11. The box-plots for the two cases are given in Figure 4 and Figure 5.

![Box-plot diagram](image_url)

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

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>BMA</th>
<th>ARMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average risk</td>
<td>0.2122</td>
<td>0.2041</td>
<td>0.1886</td>
<td>0.1856</td>
</tr>
<tr>
<td></td>
<td>(0.0111)</td>
<td>(0.0095)</td>
<td>(0.0071)</td>
<td>(0.0068)</td>
</tr>
<tr>
<td>Risk reduction</td>
<td>13%</td>
<td>9%</td>
<td>1.5%</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 10: Comparing ARMS with AIC, BIC, and BMA: Case 5.2.2

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>BMA</th>
<th>ARMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average risk</td>
<td>0.2306</td>
<td>0.2489</td>
<td>0.2289</td>
<td>0.2101</td>
</tr>
<tr>
<td></td>
<td>(0.0092)</td>
<td>(0.0108)</td>
<td>(0.0089)</td>
<td>(0.0081)</td>
</tr>
<tr>
<td>Risk reduction</td>
<td>8.5%</td>
<td>15.5%</td>
<td>8%</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 11: Comparing ARMS with AIC, BIC, and BMA: Case 5.2.3
To summarize the simulation in this subsection, both BMA and ARMS improve over model selection 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 BMA.

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 12, $CV_1$ 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 $\sigma^2$ is bigger than 0.5. All the three cross-validation methods perform worse than ARMS when $\sigma^2$ 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.

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 13, risk reduction of ARMS over the cross validation methods $CV_{68}$, $CV_{50}$, and $CV_1$ are 14%, 10%, and 15% respectively.
<table>
<thead>
<tr>
<th></th>
<th>$\sigma^2 = 0.1$</th>
<th>0.5</th>
<th>1.0</th>
<th>2.25</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV_1</td>
<td>0.0128</td>
<td>0.0614</td>
<td>0.122</td>
<td>0.256</td>
<td>0.495</td>
</tr>
<tr>
<td></td>
<td>(0.0006)</td>
<td>(0.0031)</td>
<td>(0.0056)</td>
<td>(0.012)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>AIC</td>
<td>0.0127</td>
<td>0.0614</td>
<td>0.121</td>
<td>0.254</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>(0.0006)</td>
<td>(0.0030)</td>
<td>(0.0059)</td>
<td>(0.015)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>CV_{50}</td>
<td>0.0126</td>
<td>0.0656</td>
<td>0.126</td>
<td>0.251</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>(0.0006)</td>
<td>(0.0032)</td>
<td>(0.0055)</td>
<td>(0.013)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>CV_{68}</td>
<td>0.0133</td>
<td>0.0751</td>
<td>0.127</td>
<td>0.252</td>
<td>0.472</td>
</tr>
<tr>
<td></td>
<td>(0.0007)</td>
<td>(0.0036)</td>
<td>(0.0057)</td>
<td>(0.015)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>BIC</td>
<td>0.0133</td>
<td>0.0720</td>
<td>0.128</td>
<td>0.251</td>
<td>0.479</td>
</tr>
<tr>
<td></td>
<td>(0.0007)</td>
<td>(0.0035)</td>
<td>(0.0056)</td>
<td>(0.014)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>ARMS</td>
<td>0.0135</td>
<td>0.0615</td>
<td>0.112</td>
<td>0.210</td>
<td>0.392</td>
</tr>
<tr>
<td></td>
<td>(0.0006)</td>
<td>(0.0027)</td>
<td>(0.0048)</td>
<td>(0.010)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Risk reduction</td>
<td>-6%</td>
<td>0%</td>
<td>7%</td>
<td>16%</td>
<td>18%</td>
</tr>
</tbody>
</table>

Table 12: Comparing ARMS with CV: Case 5.3.1

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>CV_{68}</th>
<th>CV_{50}</th>
<th>CV_1</th>
<th>ARMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average risk</td>
<td>0.2252</td>
<td>0.2230</td>
<td>0.2314</td>
<td>0.2219</td>
<td>0.2342</td>
<td>0.1995</td>
</tr>
<tr>
<td></td>
<td>(0.0094)</td>
<td>(0.0098)</td>
<td>(0.0134)</td>
<td>(0.0087)</td>
<td>(0.0089)</td>
<td>(0.0077)</td>
</tr>
</tbody>
</table>

Table 13: Comparing ARMS with CV: Case 5.3.2

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, at least we should be careful about the performance of the model selection methods.*

Now let’s look at two more simulation cases to check the above rule.

**Case 5.4.1.** A randomly selected model.

We randomly generated a model in the same way as in the beginning of Section 5.2, which happened to be:

$$Y = 0.35 + 1.09X_1 + 1.35X_2 + 1.11X_3 + 0.54X_4 + 0.53X_5 + 1.70X_6 + \epsilon$$

From Tables 14 and 15, ARMS performs significantly better when PIE is bigger than 0.5; model selection (at least BIC) performs better when PIE is less than 0.4. So the rule works well for this case.
Case 5.4.2 A case with 50 observations.

We consider a data set with sample size 50 instead of 100. The true model is given in (2). From Tables 16 and 17, the rule again seems to work.

6 Data examples

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

- Randomly permute the order of the observations and then split the data into two parts, with the first part ($m$ observations) used for estimation, and the second part ($n - m$ observations) as the validation set for assessment.

- Compute the predictive mean squared error (PMSE)

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

where $\hat{Y}_i$ is based on the first part of the data using the method being evaluated.

- Repeat above steps 100 times and obtain the average PMSE over the permutations.
<table>
<thead>
<tr>
<th></th>
<th>$\sigma^2 = 0.1$</th>
<th>0.5</th>
<th>1.0</th>
<th>2.25</th>
<th>4.0</th>
<th>6.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMS</td>
<td>0.0090</td>
<td>0.0521</td>
<td>0.102</td>
<td>0.221</td>
<td>0.413</td>
<td>0.578</td>
</tr>
<tr>
<td></td>
<td>(0.0005)</td>
<td>(0.0023)</td>
<td>(0.0041)</td>
<td>(0.012)</td>
<td>(0.021)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>BIC</td>
<td>0.0085</td>
<td>0.0507</td>
<td>0.109</td>
<td>0.277</td>
<td>0.560</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>(0.0006)</td>
<td>(0.0025)</td>
<td>(0.0043)</td>
<td>(0.015)</td>
<td>(0.027)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>AIC</td>
<td>0.0107</td>
<td>0.0523</td>
<td>0.105</td>
<td>0.247</td>
<td>0.492</td>
<td>0.674</td>
</tr>
<tr>
<td></td>
<td>(0.0007)</td>
<td>(0.0029)</td>
<td>(0.0051)</td>
<td>(0.014)</td>
<td>(0.027)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>Risk reduction</td>
<td>-6%</td>
<td>-3%</td>
<td>-3%</td>
<td>10%</td>
<td>16%</td>
<td>14%</td>
</tr>
</tbody>
</table>

Table 15: Risks of ARMS, AIC and BIC in Error Variance: Case 5.4.1

<table>
<thead>
<tr>
<th>$\sigma^2 = 0.01$</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>2.25</th>
<th>4.0</th>
<th>6.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>0.044</td>
<td>0.377</td>
<td>0.585</td>
<td>0.636</td>
<td>0.840</td>
<td>0.913</td>
</tr>
<tr>
<td>AIC</td>
<td>0.045</td>
<td>0.357</td>
<td>0.568</td>
<td>0.622</td>
<td>0.837</td>
<td>0.907</td>
</tr>
</tbody>
</table>

Table 16: Instability of AIC and BIC wrt Error Variance: Case 5.4.2

The results are given in Table 18. The number of explanatory variables (Column 2), split proportion $m : (n - m)$ (Column 3), and PIE (Column 4) 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 BMA 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.

3. For Geyser data C and gas vapor data D, the risks of both ARMS and BMA 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 data sets with PIE smaller than 0.5 (data sets C and D), model selection is quite stable and combining has no advantage; but for the cases with PIE larger than 0.5 (the remaining 4 data sets), model combining worked better than model selection. For the case with very high PIE (A, $B'$ and E), we see the advantage of ARMS over BMA.

7 Concluding remarks

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. Focusing on
<table>
<thead>
<tr>
<th></th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>2.25</th>
<th>4.0</th>
<th>6.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMS</td>
<td>0.00287</td>
<td>0.0266</td>
<td>0.116</td>
<td>0.215</td>
<td>0.483</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td>(0.00019)</td>
<td>(0.0012)</td>
<td>(0.0049)</td>
<td>(0.011)</td>
<td>(0.032)</td>
<td>(0.058)</td>
</tr>
<tr>
<td>BIC</td>
<td>0.00257</td>
<td>0.0292</td>
<td>0.134</td>
<td>0.246</td>
<td>0.562</td>
<td>1.292</td>
</tr>
<tr>
<td></td>
<td>(0.00015)</td>
<td>(0.0015)</td>
<td>(0.0058)</td>
<td>(0.012)</td>
<td>(0.043)</td>
<td>(0.078)</td>
</tr>
<tr>
<td>AIC</td>
<td>0.00270</td>
<td>0.0256</td>
<td>0.126</td>
<td>0.258</td>
<td>0.577</td>
<td>1.172</td>
</tr>
<tr>
<td></td>
<td>(0.00014)</td>
<td>(0.0014)</td>
<td>(0.0063)</td>
<td>(0.014)</td>
<td>(0.033)</td>
<td>(0.066)</td>
</tr>
<tr>
<td>Risk reduction</td>
<td>-11%</td>
<td>-4%</td>
<td>8%</td>
<td>13%</td>
<td>14%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Table 17: Risks of ARMS, AIC and BIC in Error Variance: Case 5.4.2

<table>
<thead>
<tr>
<th></th>
<th>d</th>
<th>Split</th>
<th>PIE</th>
<th>BMA</th>
<th>BIC</th>
<th>AIC</th>
<th>ARMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
<td>37:10</td>
<td>0.819</td>
<td>(0.0025)</td>
<td>0.0699</td>
<td>(0.0029)</td>
<td>0.0764</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9%</td>
<td>19%</td>
<td>15%</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>13</td>
<td>200:51</td>
<td>0.574</td>
<td>(0.74)</td>
<td>18.65</td>
<td>(0.87)</td>
<td>19.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2%</td>
<td>6%</td>
<td>4%</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>B'</td>
<td>13</td>
<td>50:201</td>
<td>0.729</td>
<td>(0.86)</td>
<td>22.42</td>
<td>(0.99)</td>
<td>24.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7%</td>
<td>16%</td>
<td>13%</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>180:42</td>
<td>0.317</td>
<td>(0.011)</td>
<td>0.280</td>
<td>(0.011)</td>
<td>0.279</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-1%</td>
<td>-2%</td>
<td>-2%</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>24:8</td>
<td>0.434</td>
<td>(0.37)</td>
<td>12.93</td>
<td>(0.37)</td>
<td>12.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1%</td>
<td>-2%</td>
<td>-1%</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>7</td>
<td>14:5</td>
<td>0.740</td>
<td>(0.041)</td>
<td>1.161</td>
<td>(0.055)</td>
<td>1.344</td>
</tr>
</tbody>
</table>

Table 18: Comparing ARMS with AIC, BIC and BMA on Real Data Sets

linear regression, under the global $L_2$ risk, we studied a model combining method with model screening and conducted a number of simulations for comparing model selection methods and model combining methods in representative cases and random settings. The main results of the paper are summarized below.

1. We proposed an instability index, PIE, to measure model selection instability in estimation. A rule of thumb for using the instability index 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. It was also shown that model combining can substantially reduce PIE when it is high for model selection.
2. We proposed a method for combining model, ARMS, with a step of model screening. Model screening with e.g., AIC and BIC can remove some very poor models that would hurt the combined estimator and at the same time saves computational cost.

3. Although ARMS and BMA both improve the estimation accuracy when PIE is high, the empirical comparisons show that ARMS performs better than the BMA method based on MC\(^3\) when the true model size is not very small and the error variance is not small. The risk reduction of ARMS over BMA reached 18% in one case. It seems that BMA 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.

In the current computer era, statistical tools have been increasingly used in all kinds of real world data analysis applications. Many highly sophisticated statistical procedures are now “just a click away”. In particular, automated model selection tools have been widely applied by statistical users. As statisticians, while delighted to see the increased utilization of statistical tools, we should continue to educate the users to be aware of potential dangers of indiscriminate use of such automated statistical procedures. In the meantime, to effectively prevent drawing overly-optimistic or misleading conclusions based on a single selected model, it is important for us to develop tools that properly characterize potential problems of model selection (if present) and we advocate the report of such characterizations in data analysis. In this paper, PIE was proposed from a perturbation and estimation perspective and it does seem to provide useful information regarding reliability of model selection. Other measures from different angles (e.g., prediction at a given point or parameter estimation) are also of interest. With such measures reported in data analysis, the readers can readily have a sense on how reliable the conclusions from the selected model are. We believe that more research in this area is of theoretical interest and can have an impact on real world statistical applications.

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

\[q^n_2 = \sum_{j \in \Gamma} \frac{1}{K} \Pi_{i=n_1+1}^n \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left( -\frac{1}{2\sigma_j^2} (y_i - \hat{f}_j(x_i))^2 \right).\]
Consider $\log (p^{n^2} / q^{n^2})$. Assume first that $j^* \in \Gamma_y$. By monotonicity of the log function, we have

$$
\log (p^{n^2} / q^{n^2}) \leq \log \left( \frac{\prod_{i=1}^{n} (2\pi \sigma_j^{-1})^{-1/2} \exp \left( -\frac{1}{2\sigma_j^2} (y_i - f_j(x_i))^2 \right)}{\frac{1}{N} \prod_{i=1}^{n} (2\pi \sigma_j^{-1})^{-1/2} \exp \left( -\frac{1}{2\sigma_j^2} (y_i - f_j(x_i))^2 \right)} \right)
= \log K_s + \frac{1}{2} \sum_{i=n+1}^{n} \left( \log \sigma_j^2 - \frac{(y_i - f_j(x_i))^2}{\sigma_j^2} - \frac{(y_i - f(x_i))^2}{\sigma^2} - 1 - \log \sigma^2 \right). \tag{10}
$$

Taking expectation conditional on the first part of the data, as denoted by $E_{n1}$, we have

$$
E_{n1} \left( \log \sigma_j^2 - \frac{(y_i - f_j(x_i))^2}{\sigma_j^2} - \frac{(y_i - f(x_i))^2}{\sigma^2} - 1 - \log \sigma^2 \right) = \left( \frac{\| f_j - f \|}{\sigma_j^2} + \frac{\sigma^2}{\sigma_j^2} - 1 - \log \sigma^2 \right). \tag{11}
$$

Observe that

$$
q^{n^2} = \sum_{j \in \Gamma_y} \frac{1}{\sqrt{2\pi \sigma_j^2}} \exp \left( -\frac{1}{2\sigma_j^2} (y_{n+1} - f_j(x_{n+1}))^2 \right)
\times \sum_{j \in \Gamma_y} \frac{1}{\sqrt{2\pi \sigma_j^2}} \exp \left( -\frac{1}{2\sigma_j^2} (y_{n+1} - f_j(x_{n+1}))^2 \right)
\times \sum_{j \in \Gamma_y} \frac{1}{\sqrt{2\pi \sigma_j^2}} \exp \left( -\frac{1}{2\sigma_j^2} (y_{n+1} - f_j(x_{n+1}))^2 \right)
\times \sum_{j \in \Gamma_y} \frac{1}{\sqrt{2\pi \sigma_j^2}} \exp \left( -\frac{1}{2\sigma_j^2} (y_{n+1} - f_j(x_{n+1}))^2 \right).
$$

Let $p_i = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left( -\frac{(y_i - f(x_i))^2}{2\sigma^2} \right)$ and $g_i = \sum_{j \in \Gamma_y} W_{j,i} \frac{1}{\sqrt{2\pi \sigma_j^2}} \exp \left( -\frac{(y_i - f_j(x_i))^2}{2\sigma_j^2} \right)$. It follows by the definition of $W_{j,i}$ that

$$
\log (p^{n^2} / q^{n^2}) = \sum_{i=n+1}^{n} \log \left( \frac{p_i}{g_i} \right). \tag{12}
$$

Conditioned on the first part of the data and the explanatory variables, as denoted by $E'_{n1}$ below, we have

$$
E'_{n1} \log \left( \frac{p_i}{g_i} \right) = \int p_i \log \frac{p_i}{g_i} dy_i \geq \int (\sqrt{p_i} - \sqrt{g_i})^2 dy_i.
$$

Let $\bar{f}_i(x) = \sum_{j \in \Gamma_y} W_{j,i} f_j$ for $n_1 + 1 \leq i \leq n$. By Lemma 1 of Yang (2001), under Conditions 1 and 2, we have that for $n_1 + 1 \leq i \leq n$,

$$
E'_{n1} \log \left( \frac{p_i}{g_i} \right) \geq \frac{(\bar{f}_i(x_i) - f(x_i))^2}{\sigma^2 (2(1 + \xi_2) + 9\tau)}.
$$

Together with (12), we have

$$
\sum_{i=n+1}^{n} E_{n1} \left( \frac{(\bar{f}_i(x_i) - f(x_i))^2}{\sigma^2 (2(1 + \xi_2) + 9\tau)} \right) \leq \log K_s + \frac{1}{2} n_2 \sum_{i=n+1}^{n} \left( \frac{\| f_j - f \|}{\sigma_j^2} + \frac{\sigma^2}{\sigma_j^2} - 1 - \log \sigma^2 \right). \tag{13}
$$
That is,
\[
\frac{1}{n_2} \sum_{i=n_1+1}^{n} \| \overline{f}_i - f \|^2 \leq \sigma^2 \left( 2(1 + \xi_2) + 9\tau \right) \left( \frac{\log K_s}{n} + \frac{1}{2} E_{n_1} \left( \frac{\| \hat{f}_j - f \|^2}{\sigma_j^2} + \frac{\sigma^2}{\sigma_j^2} - 1 - \log \frac{\sigma^2}{\sigma_j^2} \right) \right),
\]
By convexity of the squared $L_2$ norm, together with that $\overline{f}(x) = \frac{1}{n_2} \sum_{i=n_1+1}^{n} \overline{f}_i$, we have
\[
\| \overline{f} - f \|^2 \leq \frac{1}{n_2} \sum_{i=n_1+1}^{n} \| \overline{f}_i - f \|^2.
\]
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 $j^* \in \Gamma_s$,
\[
\| \overline{f} - f \|^2 \leq (1 + \xi_2 + 9\tau/2) \left( \frac{2\sigma^2 \log(K_s)}{n} + \frac{1}{\xi_1} \| \hat{f}_j^* - f \|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} (\sigma_j^2 - \sigma^2)^2 \right),
\]
where $C(\xi_1, \xi_2) = \frac{1}{\xi_1^{(1+\log \xi_2)}}$. Since $\overline{f}$ is a convex combination of the original estimators, under Condition 1, we have $\| \overline{f} - f \|^2 \leq \tau \sigma^2$ when $j^* \notin \Gamma_s$. Let $G_s$ denote the event that model $j^*$ is in $\Gamma_s$. It follows that
\[
\| \overline{f} - f \|^2 \leq \tau \sigma^2 I_{G_s} + (1 + \xi_2 + 9\tau/2) \left( \frac{2\sigma^2 \log(K_s)}{n} + \frac{1}{\xi_1} \| \hat{f}_j^* - f \|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} (\sigma_j^2 - \sigma^2)^2 \right) I_{G_s},
\]
where $I_{G_s}$ denote 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 contain 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)).

9 Acknowledgments

This research was supported by the US National Science Foundation CAREER Award Grant DMS094323.

References


