1999

Regression with multiple candidate models: Selecting or mixing?

Yuhong Yang
Iowa State University

Recommended Citation

Yang, Yuhong, "Regression with multiple candidate models: Selecting or mixing?" (1999). *Statistics Preprints*. 96.
http://lib.dr.iastate.edu/stat_las_preprints/96

Regression with multiple candidate models: Selecting or mixing?

Abstract

Model combining (mixing) provides an alternative to model selection. An algorithm ARM was recently proposed by the author to combine different regression models/methods. In this work, an improved risk bound for ARM is obtained. In addition to some theoretical observations on the issue of selection versus combining, simulations are conducted in the context of linear regression to compare performance of ARM with the familiar model selection criteria AIC and BIC, and also with some Bayesian model averaging (BMA) methods. The simulation suggests the following. Selection can yield a smaller risk when the random error is weak relative to the signal. However, when the random noise level gets higher, ARM produces a better or even much better estimator. That is, mixing appropriately is advantageous when there is a certain degree of uncertainty in choosing the best model. In addition, it is demonstrated that when AIC and BIC are combined, the mixed estimator automatically behaves like the better one. A comparison with bagging (Breiman (1996)) suggests that ARM does better than simply stabilizing model selection estimators. In our simulation, ARM also performs better than BMA techniques based on BIC approximation.

Keywords

ARM, combining procedures, model averaging, model selection

Disciplines

Statistics and Probability

Comments

This preprint was published as Yuhong Yang, "Regression with Multiple Candidate Models: Selecting or Mixing?", *Statistics Sinica* (2003): 783-809.

Regression with Multiple Candidate Models: Selecting or Mixing? *

Yuhong Yang
yyang@iastate.edu
Department of Statistics
Iowa State University
Ames, IA 50011

September 1, 2000

Abstract

Model averaging provides an alternative to model selection. An algorithm ARM rooted in information theory is proposed to combine different regression models/methods. A simulation is conducted in the context of linear regression to compare its performance with familiar model selection criteria AIC and BIC, and also with some Bayesian model averaging (BMA) methods.

The simulation suggests the following. Selection can yield a smaller risk when the random error is weak relative to the signal. However, when the random noise level gets higher, ARM produces a better or even much better estimator. That is, mixing is advantageous when there is a certain degree of uncertainty in choosing the right model. In addition, it is demonstrated that when AIC and BIC are combined, the mixed estimator automatically behaves like the better one. A comparison with bagging (Breiman (1996)) suggests that ARM does better than simply stablizing model selection estimators. In our simulation, ARM also performs better than BMA techniques based on BIC approximation.

ARM is a computationally feasible way to combine models and/or non-model-based procedures. It is a convex combination of the original estimators with data-dependent weights. For the determination of the weights, the data is split into two parts. The first one is used for estimation by each model or procedure and the accuracies of these estimators are assessed using the second half of the data. The accuracies are then used to assign the weights in a way such that a connection between function estimation and information theory ensures a desired theoretical capability of adaptation over different models and/or regression procedures.

Key words and phrases: ARM, adaptation, combining estimators, model averaging, model selection.

1 Introduction

In statistical applications, multiple models are often considered. Historically, one of the models is selected based on a statistical criterion together with graphical inspections. Final estimation,

*This research was partially supported by the United States National Security Agency Grant MDA9049910060.

interpretation, and prediction are then based on the selected model. Various model selection criteria have been proposed from different perspectives including minimizing the estimated prediction risk (such as AIC (Akaike (1973))), and asymptotically maximizing the posterior probability of a model (such as BIC (Schwartz (1978))) from a Bayesian's point of view. Different theoretical properties have been shown for these criteria. It is well-known that when one of the models being considered is the true model, with probability tending to 1, BIC selects the true model and it performs asymptotically better than AIC; on the other hand, if none of the models being compared is the true model, AIC asymptotically outperforms BIC in terms of statistical risks. For the reality of a finite sample, however, for either case, the answer to the question which criterion is better depends on how fast the approximation errors (bias) of the relevant models (depending on the sample size and the error variance) decrease.

Breiman (1996b) pointed out that estimators based on model selection are instable. He proposed a method *bagging* to generate multiple versions of an estimator and then average them into a stablized estimator. Empirical evidence showed advantage of bagging in terms of estimation accuracy. Another approach to reduce variability in model selection is model averaging. Bayesian model averaging is a natural way to proceed from a Bayesian point of view (see, e.g., Draper (1995) and George and McCulloch (1997)). Interesting results have been obtained on choice of priors and computation algorithms (see, e.g., Kass and Raftery (1995) and Berger and Pericchi (1996)). Some recent work has been focused on the case when a large number of models are to be combined and two methods were suggested to handle the computational difficulties that arise when summing over all the models for obtaining the posterior distribution. One approach is to restrict attention to models that are supported by the data (e.g., Madigan and Raftery (1994)) and the other uses Markov Chain Monte Marlo approximation (e.g., Madigan and York (1995)). Raftery (1995) suggests the use of BIC approximation for Bayesian model averaging. The readers are referred to a review article on this topic by Hoeting, Madigan, Raftery and Volinsky (1999) for more details. Buckland, Burnham and Augustin (1997) proposed a plausible model weighting method according to values of a model selection criterion (e.g., AIC). Cross-validation and bootstrapping have also been used to linearly combine different estimators with the intention to improve accuracy by finding the best linear combination (Wolpert (1992), Breiman (1996a), LeBlanc and Tibshirani (1996)). The objective is more aggressive than constructing an estimator to achieve the best performance among the estimators. Juditsky and Nemirovski (2000) proposed a stochastic approximation method to combine K

estimators and theoretically showed that under the squared L_2 loss, the order $(\log K)n^{-1/2}$ is basically the price one needs to pay in general for searching for the best linear combination.

Recently, combining predictors has become a very active topic in computational learning theory. The goal is to combine a set of predictors (called *experts* in the machine learning literature) appropriately so that the aggregated estimator behaves almost as well as the best predictor for all possible outcomes in terms of a cumulative loss. The emphasis is on performance bound without any probablistic assumption at all. Strategies have been proposed and shown to have this property (see, e.g., Vovk (1990), Littlestone and Warmuth (1994), Cesa-Bianchi *et al* (1997)). Adaptation results on combining density estimators in terms of statistical risks were obtained by Yang (2000a) and Catoni (1997).

In this work, a method named Adaptive Regression by Mixing (ARM) is proposed to combine different models and/or estimation procedures for regression. Our main interest in combining models is to have a smaller risk (under the square error loss) compared to model selection. ARM is derived based on earlier theoretical work in Yang (2000a) and Yang (2000b) in the context of density estimation and nonparametric regression respectively. ARM is intended to be computationally feasible while maintaining a desirable theoretical property of the theoretically attractive but impractical method in Yang (2000b). Though technically related, our motivation and the method does not follow from formal Bayesian considerations. In particular, no averaging over parameter spaces is needed. In addition, application of ARM does not require that at least one of the candidate models is correct. A simulation study is conducted in the context of linear regression. It shows the following:

1. Model selection can yield a better accuracy when a model is strongly preferred based on the data. When the noise level is higher (relative to the sample size) causing more difficulty in comparing models, ARM gives better or even much better performance.

2. Improvement of ARM over model selection goes beyond stablizing estimators based on model selection criteria.

3. When AIC and BIC are combined by ARM, it indeed performs well as if it knew which criterion is better in advance.

For simplicity in demonstration, we choose simple settings to work with in this paper. In particular, we focus on the case when the number of the competing models is not large compared to the sample size. Comparison with some Bayesian model averaging techniques in our simulation shows the advantage of ARM in terms of estimation/prediction accuracy under the square error loss. The methodology of ARM works more generally including nonparametric regression. General theoretical and simulation results on ARM focusing on nonparametric regression are in a subsequent paper (Yang (1999)).

The paper is organized as follows. In Section 2, we set up the problem of interest. In Section 3, we propose the ARM algorithms and then present simulation results in Section 4. A concluding remark is in Section 5. A theoretical property of ARM is stated in an appendix.

2 Problem of interest

Assume we observe $(Y_i, \mathbf{X}_i)$, $i = 1, ..., n$, where $\mathbf{X}_i = (X_{i1}, ..., X_{id})$ is the explanatory variable of dimension d and Y_i is the response variable. We assume that $(Y_i, \mathbf{X}_i)_{i=1}^n$ are i.i.d. copies of a random pair $(Y, \mathbf{X})$. The goal is to estimate the functional relationship between the response and the explanatory variable. Assume

$$Y = f(\mathbf{X}) + \varepsilon,$$

where $f(\mathbf{x})$ is the true underlying regression function and the random error ε is assumed to be normally distributed with unknown variance σ^2 unless stated otherwise (e.g., in Section 3.2).

To estimate f, K plausible models are being considered:

$$Y = f_k(\mathbf{x}, \theta_k) + \varepsilon,$$

where for each $k \in \{1, ..., K\}$, $\{f_k(\mathbf{x}, \theta_k),\ \theta_k \in \Theta_k\}$ is a family of regression functions with θ_k being the parameter (a vector in general). For a given model, different methods can be used to estimate θ_k. Let $\hat{\theta}_{k,n}$ be an appropriate estimator based on $Z^n = (Y_i, \mathbf{X}_i)_{i=1}^n$. Let $\hat{\sigma}_{k,n}^2$ denote an estimator of σ^2 using model k based on Z^n. For Gaussian and double-exponential errors that will be considered later, the maximum likelihood estimators will be used.

In this paper, the comparison of estimators will be focused on the statistical risk under the squared L_2 distance. Let $\hat{f}_n$ be an estimator of f based on Z^n, the risk is

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

where $P_{\mathbf{X}}$ denotes the distribution of $\mathbf{X}$ and the second expectation is taken with respect to Z^n under the true model. The risk of an estimator based on a linear model $f_k(\mathbf{x}, \theta_k)$ can be decomposed into two parts:

$$R(f, \hat{f}_{k,n}) = \int (f(\mathbf{x}) - f_k(\mathbf{x}, \theta_k^*))^2 P_{\mathbf{X}}(d\mathbf{x}) + E\int \left(f_k(\mathbf{x}, \hat{\theta}_k) - f_k(\mathbf{x}, \theta_k^*)\right)^2 P_{\mathbf{X}}(d\mathbf{x}),$$

where $f_k(\mathbf{x}, \theta_k^*)$ is the best approximator of f in the family $f_k(\mathbf{x}, \theta_k), \theta_k \in \Theta_k$, i.e., θ_k^* minimizes $\int \left(f(\mathbf{x}) - f_k(\mathbf{x}, \theta_k)\right)^2 P_{\mathbf{X}}(d\mathbf{x})$ over $\theta_k \in \Theta_k$. As the decomposition suggests, to have a small total risk, one needs a good trade-off between the approximation error (which tends to decrease as the model gets larger) and the estimation error (which tends to increase as the number of parameters increases). Even if one of the models is the true one, if it has a lot of unknown parameters relative to the number of observations, a simpler model is better in total risk when the reduction of variability in parameter estimation exceeds the bias. Of course, in applications, one does not know which models perform the best at the given sample size. The purpose of adaptive estimation is to seek an estimator whose risk is automatically close to the smallest one among the models. Model selection based on a suitable criterion is a natural way to obtain such an adaptivity. Computationally feasible adaptive estimators by combining the models (rather than selection) will be proposed and the performance will be compared with the familiar model selection criteria.

We now describe some terminology. In this paper, with the random error distribution assumed to be known up to a scale parameter, a model refers to a choice of a family of regression functions. A regression procedure (or simply a procedure) refers to a method of estimating f at each sample size. Let δ be a procedure. Given the sample size n and the observations $Z^n = (Y_i, \mathbf{X}_i)_{i=1}^n$, the procedure δ produces an estimator $\hat{f}_{\delta,n}(\mathbf{x}) = \hat{f}_{\delta,n}(\mathbf{x}; Z^n)$ of $f(\mathbf{x})$. In general, a procedure may or may not be derived based on a model.

3 Recipe for combining models and/or regression procedures

3.1 Combining models under Gaussian errors

We first propose algorithm ARM (**A**daptive **R**egression by **M**ixing) to combine multiple models under Gaussian errors. There are two main steps involved. For the first one, half of the sample is used to estimate θ_k for $1 \leq k \leq K$. At the second step, the remaining half of the sample is being predicted based on the fitted models and the predictions are assessed by comparing the predicted values with the true observations. Then the models are appropriately weighted according to the assessment of predictions in terms of a certain discrepancy measure. For simplicity, assume n is even.

Algorithm 1

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

- *Step 2.* Estimate θ_k by $\hat{\theta}_k = \hat{\theta}_{k,n/2}$ by least squares method based on $Z^{(1)}$. Find MLE of σ^2, $\hat{\sigma}_k^2 = \hat{\sigma}_{k,n/2}^2$ (again based only on $Z^{(1)}$).

- *Step 3.* Assess the accuracies of the models using the remaining half of the data $Z^{(2)}$. For each k, for $n/2+1 \le i \le n$, predict Y_i by $f_k(\mathbf{X}_i, \hat{\theta}_k)$. Compute an overall measure of discrepancy $D_k = \sum_{i=n/2+1}^{n}(Y_i - f_k(\mathbf{X}_i, \hat{\theta}_k))^2$.

- *Step 4.* Compute the weight for model k. Let
$$W_k = \frac{(\hat{\sigma}_k)^{-n/2}\exp\left(-\hat{\sigma}_k^{-2}D_k/2\right)}{\sum_{j=1}^{K}(\hat{\sigma}_j)^{-n/2}\exp\left(-\hat{\sigma}_j^{-2}D_j/2\right)}.$$
Note that $\sum_{k=1}^{K} W_k = 1$.

- *Step 5.* Compute the convex combination of the estimators produced by the models:
$$\tilde{f}_n(\mathbf{x}) = \sum_{k=1}^{K} W_k f_k(\mathbf{x}, \hat{\theta}_{k,n}).$$

Remarks:

1. If we put the uniform prior on the models and pretend that the estimates of f and σ based on the first half of the data are the true values of the models, then W_k may be interpreted as the posterior probability of model k after observing the second half of the data. Our motivation and justification, however, is not Bayesian. Our interest in combining procedures is to automatically have a small estimation/prediction risk without knowing which one works the best at the given sample size. Note that ARM is not a formal Bayes procedure. In particular, no averaging over parameters is performed. It may seem that this (approximately) corresponds to a noninformative (often improper) prior on parameters, but for comparing models, improper priors are not suitable since they do not give unique posterior model probabilities (see Berger and Pericchi (1996) for an intrinsic Bayes factor approach as a solution from a Bayesian point of view).

2. For ARM, in general, we do not require that at least one of the models is correct. The models may be only approximations as is more realistic in applications. The risk bound for ARM (as will be given in the appendix) does not need the requirement. The combined procedure performs close to the best approximating model. The BMA methods, however, assume that the models are correct (with a certain probability for each one). If one realistically regards the models as approximations, it seems unclear what the "posterior model

probabilities" really mean in the Bayesian framework. Hoeting, Madigan, Raftery and Volinsky (1999) point out that investigation when the true model is not in the candidate list is a future research direction for BMA.

3. For computing W_k, the models are treated equally with the uniform initial weight. When there are are a large number of candidate models, the uniform weighting may not be appropriate and weighting based on more subjective but reasonable considerations (e.g., more complex models receive smaller initial weights) could be applied.

Obviously the above estimator $\tilde{f}_n$ depends on the order of observation due to the partitioning. Since the observations are assumed to be independent, the order does not contain useful information for estimating f. Thus one can improve the estimator $\tilde{f}_n$ by taking the conditional expectation given the values of the observations ignoring the order. That is, in theory, one needs to compute $\tilde{f}_n$ for each permutation of the order of observations, and then average over all the permutations.

This, however, is computationally prohibitive due to the large number of permutations. Take for example $n = 50$, the total number of permutations is over 3×10^{64}. A practical solution to this difficulty is averaging over a reasonably large number of random permutations. Our experience in the simulations with the linear models as will be discussed in detail later suggests that a total number of 250 random permutation is more than sufficient there to produce very stable final estimators. Based on the above considerations, Step 5 above is replaced by the following Step $5'$.

- *Step $5'$.* Randomly permute the order of the data $(M-1)$ times. Repeat the above 4 steps and let $W_{k,r}$, $k = 1, ..., K$ denote the weight of model k computed at the r-th replication for $1 \leq r \leq M$. Let

  $$\hat{W}_k = \frac{1}{M}\sum_{r=1}^{M} W_{k,r}$$

  and let

  $$\hat{f}_n(\mathbf{x}) = \sum_{k=1}^{K} \hat{W}_k f_k(\mathbf{x}, \hat{\theta}_{k,n})$$

  be the final estimator of f. Note that it is still a convex combination of the original estimators based on the models.

3.2 Combining general regression procedures

The same idea works for combining a collection of procedures whether they are model-based or not. In addition, the Gaussian assumption on the errors can be relaxed to some extent.

Assume that the random errors ε_i's are i.i.d. with density $g(t/\sigma)/\sigma$, where $\sigma > 0$ is unknown but g is a known probability density function with respect to a measure μ with $\int tg(t)d\mu = 0$ and $0 < \int t^2 g(t)d\mu = \sigma_0^2 < \infty$. Thus the random errors have mean zero and variance $\sigma^2\sigma_0^2$. Let $\delta_1, ..., \delta_K$ be K estimation procedures with δ_j producing estimators $\hat{f}_{\delta_j,i}(\mathbf{x}) = \hat{f}_{\delta_j,i}(\mathbf{x};Z^i)$ based on observation Z^i for $i \geq 1$. An estimator $\hat{\sigma}^2_{\delta_j,i}$ is produced by the procedure based on Z^i. Some or all of the procedures could be model-based. For instance, δ_1 may be obtained based on a linear family $f(\mathbf{x},\theta)$. Then $\hat{f}_{\delta_1,i}(\mathbf{x}) = f(\mathbf{x}, \hat{\theta}_i)$ with $\hat{\theta}_i$ appropriately estimated based on the new assumption on the errors. Another procedure, say, δ_2, may be based on a nearest neighbor rule. Though the algorithm does not require that the procedures to be combined are based on the assumption on the errors, the final adaptive estimator can not behave well unless there is at least one procedure that works well for the true model. The procedures are allowed to share variance estimators if desired. The adaptation algorithm ARM works as follows.

Algorithm 2

- *Steps 0-2.* As before.
- *Step 3.* For each j, evaluate predictions. For $n/2+1 \leq i \leq n$, predict Y_i by $\hat{f}_{\delta_j,n/2}(\mathbf{X}_i)$. Compute
$$E_j = \left(\hat{\sigma}_{\delta_j,n/2}\right)^{-n/2} \Pi_{i=n/2+1}^{n} g\left(\frac{(Y_i - \hat{f}_{\delta_j,n/2}(\mathbf{X}_i))}{\hat{\sigma}_{\delta_j,n/2}}\right).$$
- *Step 4.* Compute the weight for model j. Let
$$W_j = \frac{E_j}{\sum_{l=1}^{K} E_l}.$$
- *Step 5.* Repeat steps 0-4 and average the weights over the random permutations and obtain the final estimator as before.

A particular choice of g, namely the double-exponential density, is of special interest.

Example 1: Double-exponential errors. Consider

$$g(t) = 0.5e^{-|t|}, \quad t \in R.$$

For a parametric model $f_k(\mathbf{x}, \theta_\mathbf{k})$, based on Z^n, the maximum likelihood estimator of θ_k minimizes $\sum_{i=1}^n |Y_i - f_k(\mathbf{x}, \theta_\mathbf{k})|$ and σ is estimated by $\hat{\sigma} = (1/n) \sum_{i=1}^n |Y_i - f_k(\mathbf{x}, \hat{\theta}_\mathbf{k,n})|$. The computation of the estimators can be carried out through linear programming. This is the familiar L_1 regression as widely considered for robust estimation.

Due to the use of a number of replications in the above 5 steps, the algorithm ARM can be computationally intensive if M is large. If the original estimators are computationally feasible, so is ARM as long as that there are not too many models/procedures to be combined.

3.3 Combining AIC and BIC as an illustration of ARM

As is well-known, neither AIC nor BIC performs better all the time. Roughly speaking, AIC performs better when the approximation errors of the good competing models (relative to the sampler size and σ^2) decrease slowly. An interesting question then is, can the strengths of AIC and BIC be combined?

This question has been previously considered theoretically. Barron, Yang, and Yu (1994) showed in theory that a suitable minimum description length (MDL) criterion for function estimation automatically behaves like AIC or BIC when AIC or BIC works better. The resulting estimator then is optimal both for some parametric families and also for nonparametric classes. More recently, Yu and Hansen (1999) proposes a different MDL criterion to bridge AIC and BIC. They showed in theory the procedure is both consistent (as BIC) and asymptotically optimal (as AIC).

The algorithm ARM can be directly used to combine AIC and BIC practically in the hope that it will work well regardless of which one is better. Assume, for example, Gaussian errors and consider parametric families $f_k(\mathbf{x}, \theta_k)$, $k = 1, ..., K$. Let $\hat{\theta}_{k,i}$ and $\hat{\sigma}_{k,i}$ be MLE of θ_k and σ respectively based on Z^i, $i \geq 1$. Let $\hat{k}_{AIC,i}$ and $\hat{k}_{BIC,i}$ be the model selected by AIC and BIC respectively at the given sample size. Then the procedure AIC produces an estimator $f_{\hat{k}_{AIC,i}}(\mathbf{x}, \hat{\theta}_{\hat{k}_{AIC,i}})$ of f and $\hat{\sigma}_{\hat{k}_{AIC,i},i}$ of σ based on Z^i. Similarly define the estimators based on BIC. Then the two procedures can be combined using Algorithm 2.

4 A simulation study

We consider several simple settings in the simulation to illustrate the advantage of ARM. The study was carried out using Splus.

4.1 Two non-nested models

Consider two non-nested models both with three unknown parameters:

$$\begin{aligned} Y_i &= \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i, \\ Y_i &= \theta_1 X_{2i} + \theta_2 X_{3i} + \theta_3 X_{4i} + \varepsilon_i. \end{aligned}$$

The explanatory variables X_{1i}, X_{2i}, X_{3i} are generated independently according to the uniform distribution on $[0, 1]$. The other variable X_{4i} is generated as $0.25X_{1i} + 0.75X_{5i}$, where X_{5i} is also uniformly distributed independent of X_{1i}, X_{2i}, and X_{3i}. This way X_{1i} and X_{4i} are somewhat correlated, which may be more realistic in some applications (in fact, a simulation under independence between X_{1i} and X_{4i} gave very similar conclusions). The first model is used to generate the responses with true parameters $\beta_1 = 1.0$, $\beta_2 = 0.8$, $\beta_3 = 0.9$ and with Gaussian errors. The sample size is fixed at $n = 50$ and M is taken to be 250 for ARM. Several noise levels are considered for the comparison of selection and ARM. For this case, since the two models have the same number of parameters, AIC and BIC are equivalent and just select the model with smaller residual sum of squares. The squared L_2 losses of the estimators are simulated as the average of the squared differences between the true regression function and the estimator at 500 new design points independently generated according to the same distribution. There are 200 replications and the losses are averaged over the replications to approximate the true risks of the estimators. The numbers of times (out of 200 replications) that the selection criterion chose a wrong model are also given. The numbers in the parenthesis are the corresponding standard errors.

	$\sigma^2 = 0.1$	0.3	0.5	1.0	1.5	2.0	3.0
Risk of Selection	0.00065 (4.1×10^{-5})	0.0056 (0.0003)	0.0167 (0.0011)	0.0770 (0.0051)	0.1781 (0.0093)	0.2906 (0.0164)	0.6728 (0.0359)
Mis-selection Times	0	0	1	26	57	61	81
Risk of ARM	0.00065 (4.1×10^{-5})	0.0064 (0.0004)	0.0204 (0.0012)	0.0720 (0.0040)	0.1566 (0.0080)	0.2444 (0.0134)	0.5424 (0.0321)

Table 1: Comparing Selection and Mixing for Non-nested Models

From the table, when $\sigma^2 \geq 1.0$, ARM produces a smaller risk than that based on model selection. The risk reductions are 6%, 12%, 16%, and 20% respectively, and they are extremely significant by a large sample two-sided test for paired data (the p-values are well below 0.0001) except for $\sigma^2 = 1.0$ (for which case the p-value is 0.016). Note that when $\sigma^2 \leq 0.5$, the model selection criterion basically has no difficulty finding the right model. For such a case, mixing with the wrong model can hurt the performance. Indeed, for $\sigma^2 = 0.3$ and 0.5, ARM increases the risk by 14% and 22% respectively. When σ^2 gets larger, the chance of selecting a bad model is no longer negligible and it increases the variability of the estimator based on selection. In contrast, ARM does a better job by reducing the variability in estimation through appropriate mixing instead of selecting.

It is worth pointing out that several values of σ^2 smaller than 0.1 were also considered and the risks of selection and ARM are pretty much the same, suggesting insignificant difference between selection and ARM when σ^2 is really small.

4.2 Nested models

Consider five nested models: for $1 \leq i \leq n$,

$$\begin{aligned} Y_i &= \beta_1 X_{1i} + \varepsilon_i, \\ &\quad \ldots \\ Y_i &= \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i. \end{aligned}$$

The explanatory variables are generated independently with uniform distribution on [0,1]. The errors are assumed to be independent and normally distributed with unknown variance σ^2.

This simulation addresses several issues: comparison among AIC, BIC and ARM, and combining AIC and BIC as discussed in the previous section. As mentioned in the introduction, bagging instable estimators can improve accuracy dramatically (Breiman (1996b)). It is thus of interest to compare the improvement of ARM over AIC and BIC to that by the method of bagging.

For Table 2, the data is generated according to the fourth model above with true parameters $\beta_1 = 1.0$, $\beta_2 = 0.9$, $\beta_3 = 0.8$, and $\beta_4 = 0.6$. For Table 3, the true model is the second one above with true parameters $\beta_1 = 1.0$, $\beta_2 = 0.9$. At the chosen sample size $n = 50$, AIC works better for the first case (except for $\sigma^2 = 0.1$) and BIC works better for the second. The numbers of wrong selections (out of 200 replications) are also given. The number of permutations, M, for ARM and the number of bootstrap samples for bagging are both taken to be 300.

The findings are summarized as follows.

1. For both cases, when $\sigma^2 \geq 0.5$, bagging reduces the risk of BIC significantly even up to 33% and 21% respectively. For AIC, however, bagging actually increases the risk (quite

	$\sigma^2 = 0.1$	0.5	1.0	1.5	2.0	3.0
Risk of AIC	0.0101 (0.0006)	0.0590 (0.0026)	0.1174 (0.0051)	0.1734 (0.0083)	0.2243 (0.0102)	0.3262 (0.0152)
Mis-selection by AIC	36	84	139	145	157	169
Risk of AIC$_{\text{bag}}$	0.0104 (0.0005)	0.0546 (0.0024)	0.1007 (0.0044)	0.1548 (0.0078)	0.1869 (0.0089)	0.2835 (0.0143)
Risk of BIC	0.0099 (0.0006)	0.0675 (0.0032)	0.1407 (0.0052)	0.2079 (0.0095)	0.2651 (0.0105)	0.3680 (0.0134)
Mis-selection by BIC	14	102	168	169	182	190
Risk of BIC$_{\text{bag}}$	0.0100 (0.0005)	0.0547 (0.0025)	0.1021 (0.0041)	0.1504 (0.0075)	0.1770 (0.0079)	0.2638 (0.0117)
Risk of ARM	0.0118 (0.0006)	0.0524 (0.0026)	0.0970 (0.0040)	0.1351 (0.0064)	0.1574 (0.0071)	0.2275 (0.0099)
AIC-BIC Combined	0.0097 (0.0006)	0.05960 (0.0027)	0.1184 (0.0049)	0.1749 (0.0083)	0.2171 (0.0094)	0.3161 (0.0135)

Table 2: Selection vs Mixing When the True Model Has 4 Terms

substantially when σ^2 is small) for the second case. Note that though AIC outperforms BIC for the first case (except when $\sigma^2 = 0.1$), bagging improves BIC more than AIC and makes BIC$_{\text{bag}}$ better than AIC$_{\text{bag}}$ for larger σ^2. We do not have a good explanation for these phenomena.

2. ARM works better or much better than both AIC, BIC and their bagging versions when $\sigma^2 > 0.1$. In fact, the risk of ARM here is smaller than the best among the four estimators. For the first case above, the corresponding percentages of risk reduction over the best of AIC, BIC, AIC$_{\text{bag}}$ and BIC$_{\text{bag}}$ are tabled as follows:

	$\sigma^2 = 0.5$	1.0	1.5	2.0	3.0
Risk Reduction	4%	5%	10%	11%	14%

For the second case, the reduction rates are

	$\sigma^2 = 0.5$	1.0	1.5	2.0	3.0
Risk Reduction	7%	17%	18%	16%	21%

3. When AIC and BIC are combined by ARM, the estimator automatically behaves like the better one of AIC and BIC as intended.

4.3 L_1-regression

Given in Table 4 is the result of a simulation to compare the performance of model selections and ARM under the double-exponential errors. The first 4 models considered in the previous

	$\sigma^2 = 0.1$	0.5	1.0	1.5	2.0	3.0
Risk of AIC	0.0069 (0.0005)	0.0376 (0.0028)	0.0747 (0.0048)	0.1303 (0.0087)	0.1477 (0.0091)	0.2322 (0.0144)
Mis-selection by AIC	49	59	69	97	105	129
Risk of AIC_{bag}	0.0082 (0.0005)	0.0430 (0.0024)	0.0815 (0.0042)	0.1343 (0.0079)	0.1554 (0.0086)	0.2441 (0.0137)
Risk of BIC	0.0051 (0.0004)	0.0361 (0.0031)	0.0771 (0.0044)	0.1271 (0.0072)	0.1428 (0.0066)	0.2079 (0.0115)
Mis-selection by BIC	18	35	74	113	134	142
Risk of BIC_{bag}	0.0058 (0.0004)	0.0344 (0.0021)	0.0625 (0.0035)	0.1001 (0.0058)	0.1157 (0.0069)	0.1713 (0.0107)
Risk of ARM	0.0057 (0.0003)	0.0321 (0.0019)	0.0520 (0.0030)	0.0825 (0.0045)	0.0967 (0.0058)	0.1353 (0.0079)
AIC-BIC Combined	0.0058 (0.0005)	0.0345 (0.0025)	0.0686 (0.0040)	0.1193 (0.0075)	0.1343 (0.0075)	0.2099 (0.0125)

Table 3: Selection vs Mixing When the True Model Has 2 Terms

subsection are considered here. The correct model is chosen to be the second one with true parameters $\beta_1 = 1.0$ and $\beta_2 = 0.9$. The four explanatory variables are chosen to be i.i.d. with uniform distribution on $[0, 1]$. We fix the sample size to be $n = 50$ and $M = 200$.

	$\sigma = 0.2$	0.4	0.6	0.8	1.0	1.5	3.0
Risk of AIC	0.0037 (0.0003)	0.0176 (0.0014)	0.0362 (0.0033)	0.0619 (0.0047)	0.1084 (0.0087)	0.2404 (0.0181)	0.7962 (0.0688)
Mis-selection by AIC	50	58	60	76	88	120	157
Risk of BIC	0.0029 (0.0003)	0.0141 (0.0014)	0.0300 (0.0030)	0.0582 (0.0045)	0.1029 (0.0083)	0.1807 (0.0120)	0.5212 (0.0485)
Mis-selection by BIC	16	21	27	53	78	132	176
Risk of ARM	0.0028 (0.0002)	0.0141 (0.0011)	0.0252 (0.0017)	0.0447 (0.0031)	0.0648 (0.0051)	0.1208 (0.0080)	0.4637 (0.0336)

Table 4: Selection vs Mixing with Double-Exponential Errors

The advantage of ARM is clearly seen from the simulation. Even when σ is as small as 0.2 and 0.4, ARM performs as well as BIC (AIC is significantly worse). The risk reduction rates for $\sigma \geq 0.6$ are

	$\sigma = 0.6$	0.8	1.0	1.5	3.0
Risk Reduction	16%	23%	38%	33%	11%

When σ^2 is really small (other values such as 0.05 and 0.01 not given in the above table were also considered), no significant differences were found between ARM and BIC (for this scenario, BIC should perform better than AIC).

4.4 Combining subset models when the number of predictors is small

Suppose there are 4 independent predictors uniformly distributed in [0,1]. Consider all subset models and use ARM and some BMA techniques to combine them. The true mode is one of the following:

$$\begin{aligned}
\text{Case 1} &: \quad Y = 1 + X_1 + \varepsilon, \\
\text{Case 2} &: \quad Y = 1 + X_1 + X_2 + \varepsilon, \\
\text{Case 3} &: \quad Y = 1 + X_1 + X_2 + X_3 + \varepsilon, \\
\text{Case 4} &: \quad Y = 1 + X_1 + X_2 + X_3 + X_4 + \varepsilon,
\end{aligned}$$

where the error ε has a standard normal distribution (the variance of ε is unknown to the estimators). The sample size is 50. The number of permutations for ARM is 50. The chosen BMA program for comparison, `bicreg` in Splus based on BIC approximation, was written by Adrian Raftery and revised by Chris Volinsky (available at `http://www.research.att.com/ volinsky/bma.html`). In addition to computing the posterior probabilities of all the models, one option is provided in `bicreg` to remove some unlikely models based on Occam's Window and return a more parsimonious list of models (see Raftery (1995)) (the corresponding estimator is denoted BMA_{OW}).

The squared L_2 risk of the regression estimators based on ARM and BMA are summarized in Table 5 based on 100 runs.

	Case 1	Case 2	Case 3	Case 4
BMA_{OW}	0.0853 (0.0060)	0.1182 (0.0058)	0.1781 (0.0079)	0.2326 (0.0087)
BMA	0.0738 (0.0054)	0.1015 (0.0054)	0.1443 (0.0065)	0.1837 (0.0074)
ARM	0.0706 (0.0052)	0.0789 (0.0046)	0.1093 (0.0062)	0.1312 (0.0060)

Table 5: Comparing ARM with BMA

In this simulation, ARM does substantially better compared to BMA for all of the 4 cases. The risk improvement is 4%, 22%, 24%, 29% respectively.

4.5 Crime data

Consider a crime data studied via Bayesian model averaging for illustration by e.g., Raftery, Madigan and Hoeting (1999), Fernández, Ley and Steel (1999) and originally by Ehrlich (1973).

The data contain information of 47 states in US. The response variable is the crime rate and there are 15 candidate predictors. As in those work, log-transformation of Y and the predictors were applied and linear models are considered.

Raftery et al (1999) showed that BMA gives better predictive performance compared to model selection based on Efroymson's stepwise method (Miller (1990)) and other criteria. Predictive coverage was computed based on randomly splitting the data into training and trest sets. Their analysis showed that the predictive coverage intended at 90% were about 80% for the BMA methods and were between 58% and 67% for model selection methods. While all performing poorly, BMA did significantly better.

We here compare predictive performance of ARM with BMA and Efroymson's method in terms of predictive mean squared error (PMSE). We randomly select 37 states as the training set and the remaining 10 states form the test set for computing PMSE. For ARM, differently from combining all subset models (which would be too time-consuming for ARM in the current form), here we combine some plausible models to reduce computational cost. The stepwise (forward) selection method is used to order the predictors according to the order of appearance. Then the corresponding nested model are combined with ARM. Table 6 summarizes the PMSE's of the different methods based on 200 independent runs.

	BMA_{OW}	BMA	Efroymson	ARM
PMSE	0.0736 (0.0020)	0.0702 (0.0019)	0.0746 (0.0020)	0.0659 (0.0019)

Table 6: Comparing BMA, Efroymson's Method and ARM on a Crime Data

From the table, the BMA method without using the Occam's Window does significantly better than model selection by Efroymson's method, but ARM further improves the prediction accuracy by about 6%.

5 Conclusion and discussion

A computationally feasible algorithm ARM is proposed to combine different regression models and/or regression procedures to obtain capability of adaptation. The simulation results clearly suggest advantage of ARM in terms of squared L_2 risk over the popular model selection criteria AIC and BIC when the random noise reaches a certain level. The reduction of the risk over the better one of AIC and BIC can be nearly 40%. Comparison with bagging suggests the advantage of ARM goes beyond simply stablizing the estimators based on model selection. When the error variance is smaller so that there is not much difficulty comparing the models, AIC and BIC can outperforms ARM (for this case, bagging can also increase the risk). When the error variance is really small, ARM and BIC behave equally well. The simulation also supports that when AIC and BIC are combined by ARM, the new estimator automatically behaves like the better criterion of AIC and BIC in terms of the statistical risks.

Model selection can be viewed as a model averaging with a degenerate weight distribution. Intuitively, it seems clear that when two models are hard to be distinguished at a given sample size, compared to averaging the models, selection can bring in much larger variability in the estimator. On the other hand, when one model is clearly inferior based on the data, averaging with it can damage the performance unless its assigned weight is small enough (which seems to happen with ARM when σ^2 is really small). This roughly explains the difference between selection and mixing.

For applications, one does not know before hand if selection or mixing is better. It is tempting to construct an estimation procedure that automatic switches between selection and mixing to enjoy the advantages of both schemes. So far, our several attempts did not give us the desired simulation results.

The method of bagging has been suggested to reduce the variability in model selection. Our experiments showed that bagging can have quite different effects on AIC and BIC: it consistently reduced risk for BIC when σ^2 is not too small but it hurt AIC for one case at all levels of σ^2 being considered. Further understanding of when bagging works will be very helpful. The simulations showed that ARM consistently performed better than bagging AIC and BIC.

Bayesian model averaging techniques have been proposed mainly under normal errors, some of which intend to handle cases when there are a large number of candidate models. In general, posterior model probabilities are very sensitive to the specification of the priors (cf. Fernández, Ley and Steel (1999)). For the ARM procedure in this paper, we focused on the situation with a small or moderate number of competing models. The simulation results showed advantage of ARM over BMA methods based on BIC approximation in terms of estimation/prediction accuracy under the squared error loss. To deal with a lot of candidate models, ARM needs to be speeded up, and perhaps non-uniform initial weights on models could be incorporated to regulate the comparison in terms of e.g. model complexities. It is of interest to study the actual performance and compare it with other related techniques including BMA.

Methods have been proposed to linearly combine estimators (not necessarily requiring convexity) by Wolpert (1992), Breiman (1996a), LeBlanc and Tibshirani (1996) and Juditsky and Nemirovski (2000). The objective of these methods is more aggressive than that of ARM since they try to get the best linear combination of the original estimators while ARM tries to achieve the best performance among the estimators. Theoretically speaking, one would expect to pay a higher price for searching for the best linear combination. Indeed, under the squared L_2 loss, Juditsky and Nemirovski (2000) showed that with K estimators to be combined, in general no method can result in a risk within a smaller order than $(\log K/n)^{1/2}$ from the risk of the best linear combination. For ARM, however, as will be seen next in the appendix, the convergence rate is well within order $\log K/n$ from the best estimator being considered. It is of interest to compare these methods empirically in the future.

6 Appendix: A theoretic property of ARM

The ARM algorithm in this paper is based on earlier work on combining different regression procedures in Yang (2000b). It has a close relationship to some work on data compression in information theory (see, Barron (1987), Barron and Cover (1991), Yang and Barron (1999), and others). Differently from the mainly theoretical work in Yang (2000b), ARM in this paper is designed to be computationally feasible for applications. In the determination of weights of the models/procedures in Yang (2000b), sequential predictions of one observation after another are performed and then the discrepancies in prediction are assessed for assigning the weights. These sequential computations are accordingly very costly. For the ARM algorithms in this paper, the data are split into two parts, with part one used for estimation based on each of the procedures being considered, and the predictions are then made for the second half of the data using each of the estimates (without updating the estimates after each additional observation). This in general hurts the risk bound slightly in theory, though it typically does not damage the rate of convergence. The computational gain can be substantial when the sample size is large.

As in Algorithm 2, let the first half of the data be used to estimate the regression function by each procedure. Then for each $n/2+1 \leq i \leq n$, compute the weights for the procedures similarly as in Algorithm 2 but only use the observations up till i (instead of n). Denote the weights by $W_{j,i}$ for $1 \leq j \leq K$. Let $\hat{f}_{\delta_j,n/2}(\mathbf{x};Z^{(1)})$ be the estimator based on the first half of the data and let $\hat{f}_{\Delta,i}(\mathbf{x};Z^{(1)}) = \sum_j W_{j,i}\hat{f}_{\delta_j,n/2}$ be the combined estimator based on Z^i for $n/2+1 \leq i \leq n$. Then the average cumulative risk of these estimators of f at the sample sizes between $n/2+1$ and n is

$$R_{seq}(\Delta; n) = \frac{1}{(n/2)} \sum_{i=n/2+1}^{n} E \parallel f - \hat{f}_{\Delta,i} \parallel^2 .$$

Assume that the true regression function is bounded, i.e., $\parallel f \parallel_\infty = c_f < \infty$. For simplicity, assume the errors are normally distributed with known variance. A similar result in much greater generality (without normality and with unknown variance) is in a subsequent paper focusing on nonparametric estimation (Yang (1999)).

Proposition 1: The average cumulative risk of the combined strategy is upper bounded as follows:

$$R_{seq}(\Delta; n) \leq C \inf_{1 \leq j \leq K} \left(\frac{n-2}{n} E \parallel f - \hat{f}_{\delta_j, n/2} \parallel^2 + \frac{2}{n} E \parallel f - \hat{f}_{\delta_j, n} \parallel^2 + \frac{\log K}{n} \right),$$

where the constant C depends only on c_f and σ^2.

Remarks:

1. The risk bound suggests that the adaptive estimator ARM is insensitive to addition of some bad procedures because the bound does not depend on the risks of bad estimators. Thus in some sense we do not need to worry much about ruining the final estimator when considering some procedures that are risky but with potential gains. Of course, the performance of ARM can be severely damaged if a lot of poor procedures are present (for which case the term $\log K/n$ is not small).

2. As in Yang (2000b), an estimator can be constructed to have a good risk bound at a given sample size instead of in terms of the average cumulative risk. The estimator, however, is more costly in computation.

For typical applications, $E \parallel f - \hat{f}_{\delta_j, n/2} \parallel^2$ and $E \parallel f - \hat{f}_{\delta_j, n} \parallel^2$ are of the same order. Then we have $R_{seq}(\Delta; n) \leq \tilde{C} \inf_{1 \leq j \leq K} \left(E \parallel f - \hat{f}_{\delta_j, n} \parallel^2 + \frac{\log K}{n} \right)$ for some constant $\tilde{C}$ (depending on σ^2, f, and possibly on the procedures, but not on n). Since $(\log K)/n$ does not affect the rate of convergence of the risk, the proposition then implies that the combined estimator based on ARM converges automatically at the best rate offered by the procedures for the unknown regression function. This is very useful for nonparametric regression where various rates of convergence are possible with different nonparametric procedures and for different target regression classes. For parametric regression, though the theory above does not guarantee advantage of ARM beyond rate of convergence, the simulations do strongly suggest the effectiveness of ARM in accuracy.

7 Acknowledgments

The author thanks two anonymous reviewers for their very insightful comments, which substantially improved the quality of the paper.

References

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

[2] Barron, A.R. (1987) Are Bayes rules consistent in information? In *Open Problems in Communication and Computation*, pp. 85-91. T. M. Cover and B. Gopinath editors, Springer-Verlag.

[3] Barron, A.R., and Cover, T.M. (1991) Minimum complexity density estimation. *IEEE, Trans. on Information Theory*, **37**, 1034-1054.

[4] Barron, A.R., Yang, Y., and Yu, B. (1994) Asymptotically optimal function estimation by minimum complexity criteria. In *Proc. 1994 Int. Symp. Info. Theory*, p. 38, Trondheim, Norway.

[5] Berger, J.O., and Pericchi, L.R. (1996) The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.*, **91**, 109-122.

[6] Breiman, L. (1996a) Stacked regressions. *Machine Learning*, **24**, 49-64.

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

[8] Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1995) Model selection: An integral part of inference. *Biometrics*, **53**, 603-618.

[9] Catoni, O. (1997) The mixture approach to universal model selection. Technical Report LIENS-97-22, Ecole Normale Superieure, Paris, France.

[10] Cesa-Bianchi, N., Freund, Y., Haussler, D. P., Schapire, R. and Warmuth, M. K. (1997) How to use expert advise? *Journal of the ACM* **44**, 427-485.

[11] Draper, D. (1995) Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B*, **57**, 45-97.

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

[13] Fernández, C., Ley, E. and Steel, M. F. J. (1998) Benchmark Priors For Bayesian Model Averaging. Documento de Trabajo 98-06, Fedea, Madrid, Spain.

[14] George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339-373.

[15] Hansen, M., and Yu, B. (1999) Bridging AIC and BIC: An MDL model selection criterion. In *Proceedings of IEEE Information Theory Workshop on Detection, Estimation, Classification and Imaging*, Santa Fe, NM, p 63.

[16] Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: A tutotial. *Statistical Science*, **14**.

[17] Juditsky, A. and Nemirovski, A. (2000) Functional aggregation for nonparametric estimation. To appear in *Ann. Statistics.*

[18] Kass, R.E., and Raftery, A.E. (1995) Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773-795.

[19] LeBlanc, M. and Tibshirani, R (1996) Combining estimates in regression and classification. *J. Amer. Statist. Asso.*, **91**, 1641-1650.

[20] Littlestone, N. and Warmuth, M.K. (1994) The weighted majority algorithm. *Information and Computation* **108**, 212-261.

[21] Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Asso.*, **89**, 1535-1546.

[22] Miller, A. J. (1990) *Subset Selection in Regression*, New York, Chapman-Hall.

[23] Madigan, D. and York, J. (1995) Bayesian graphical models for discreate data. *Int. Statist. Rev.*, **63**, 215-232.

[24] Raftery, A. E. (1995) Bayesian model selection in social research (with Discussion). In *Sociological Methodology* (Peter V. Marsden, ed.), 111-196, Cambridge, Mass.: Blackwells.

[25] Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian Model Averaging for Linear Regression Models. *J. Amer. Statist. Asso.*, **92**, 179-191.

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

[27] Vovk, V.G. (1990) Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pp. 372-383, 1990.

[28] Yang, Y. (1999) Adaptive Regression by Mixing. Technical Report # 12, Department of Statistics, Iowa State University.

[29] Yang, Y. (2000a) Mixing strategies for density estimation. To appear in *Ann. Statistics.*

[30] Yang, Y. (2000b) Combining different regression procedures for adaptive regression. *Journal of Multivariate Analysis*, **74**, 135-161.

[31] Yang, Y. and Barron, A.R. (1999) Information-theoretic determination of minimax rates of convergence. *Ann. Statistics*, **27**, 1564-1599.