Statistica Sinica **13**(2003), 783-809

REGRESSION WITH MULTIPLE CANDIDATE MODELS: SELECTING OR MIXING?

Yuhong Yang

Iowa State University

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.

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

1. Introduction

In statistical applications, multiple models are often considered. Historically, one of the models is selected based on hypothesis testing or the use of a statistical criterion together with graphical inspections. Final estimation, 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 (Schwarz (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; 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 unstable. He proposed a method *bagging* to generate multiple bootstrap versions of an estimator and then average them into a stabilized 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 Carlo 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.

In this paper, our interest is on the estimation of the regression function under a global performance measure. The goal is two-fold. First, we derive a performance bound on a model/procedure combining method, named ARM (adaptive regression by mixing) and compare its performance with some Bayesian model averaging techniques in some simulations; second, we study the relationship between selection and combining. We give some theoretical observations and compare performances of the combining method ARM with model selection and related methods (e.g., bagging) in simulations.

ARM was proposed in Yang (2001a) and was applied to combine nonparametric regression methods. The risk bound derived in this paper significantly improves over the earlier one there. Unlike Yang (2001a), parametric settings are considered in this work for the purpose of studying the issue of combining versus selection.

The paper is organized as follows. In Section 2, we set up the problem of interest. In Section 3, we present the ARM algorithms and give a risk bound to theoretically characterize its performance. Section 4 addresses the issue of combining versus selection from a theoretical point of view. Intensive simulation results are given in Section 5. Concluding remarks are in Section 6. The proofs of the main theoretical results are given in an appendix.

2. Problem Setup

Assume we observe $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ 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 $(Y, \mathbf{X})$. Such a setting is commonly used in regression with a random design (see e.g., Efromovich (1999), Section 4.1). 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 independent of $\mathbf{X}$ and 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 $k \in \{1, \ldots, K\}$, $\{f_k(\mathbf{x}, \theta_k),\ \theta_k \in \Theta_k\}$ 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 $\hat{\sigma}^2_{k,n}$ denote an estimator of σ^2 using model k based on Z^n. For Gaussian and double-exponential errors considered later, maximum likelihood estimators will be used.

In this paper, the comparison of estimators will be focused on the statistical risk under squared L_2 distance. For $\hat{f}_n$ 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 approximation of f in the family $f_k(\mathbf{x}, \theta_k), \theta_k \in \Theta_k$, i.e., θ_k^* minimizes $\int (f(\mathbf{x}) - f_k(\mathbf{x}, \theta_k))^2 P_{\mathbf{X}}(d\mathbf{x})$ 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). 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 adaptivity. Alternatively, computationally feasible adaptive estimators by combining the models (rather than selection) will be presented.

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. Combining Models and/or Regression Procedures by ARM

3.1. Combining models under Gaussian errors

An algorithm ARM (**a**daptive **r**egression by **m**ixing) was proposed in Yang (2001a) to combine multiple models. There are two main steps involved. For the first one, half of the sample is used to estimate θ_k for $1 \le k \le K$. At the second step, the remaining half of the sample is predicted based on the fitted models and predictions are assessed by comparing predicted values with observations. Then the models are appropriately weighted according to the assessment of predictions provided by a 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 a least squares method based on $Z^{(1)}$. Find, e.g., the 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 the 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

$$W_k = \frac{(\hat{\sigma}_k)^{-n/2} \exp\left(-\hat{\sigma}_k^{-2} D_k/2\right)}{\sum_{j=1}^{K} (\hat{\sigma}_j)^{-n/2} \exp\left(-\hat{\sigma}_j^{-2} D_j/2\right)}$$

  for model k. Note that $\sum_{k=1}^{K} W_k = 1$.

- *Step* 5. Compute the convex combination of the estimators produced by the models:

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

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 corresponds (approximately) to a non-informative, 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. Note that $\widetilde{f}_n$ depends on all the estimators from the candidate models. This causes a difficulty in interpretation. Model selection, on the other hand, has a potential dimension reduction feature for interpreting the relationship between the response and the explanatory variables.
3. For computing W_k, models are given uniform initial weight. When there are a large number of candidate models, 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 $\widetilde{f}_n$ 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 $\widetilde{f}_n$ by taking the conditional expectation given the values of the observations ignoring the order. That is, in theory, one needs to compute $\widetilde{f}_n$ 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. A practical solution to this difficulty is averaging over a reasonably large number of random permutations. Our experience in simulations with linear models, as will be discussed in detail later, suggests that a total of 250 random permutation is more than sufficient 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 four steps and let $W_{k,r}$, $k=1,\ldots,K$ denote the weight of model k computed at the r-th time for $1 \le r \le M$. Let $\hat{W}_k = 1/M \sum_{r=1}^{M} W_{k,r}$ 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 $\delta_1, \ldots, \delta_K$ be K estimation procedures with δ_j producing estimators $\hat{f}_{j,i}(\mathbf{x}) = \hat{f}_{j,i}(\mathbf{x};Z^i)$ based on observation Z^i for $i \ge 1$. An estimator $\hat{\sigma}^2_{j,i}$ 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 $\hat{f}_{j,i}(\mathbf{x}) = f(\mathbf{x}, \hat{\theta}_i)$ 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 be based on the assumption on the errors, the final adaptive estimator cannot 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 computation of weights is modified as follows.

For each j, for $n/2+1 \le i \le n$, predict Y_i by $\hat{f}_{j,n/2}(\mathbf{X}_i)$. Compute

$$E_j = \left(\hat{\sigma}_{j,n/2}\right)^{-n/2} \Pi_{i=n/2+1}^{n} g\left(\frac{(Y_i - \hat{f}_{j,n/2}(\mathbf{X}_i))}{\hat{\sigma}_{j,n/2}}\right).$$

Then define weight $W_j = E_j / \sum_{l=1}^{K} E_l$. As before, one can average the weights over a number of random permutations of the data to reduce the dependence of the final estimator on the order of the data.

A particular choice of g, namely the double-exponential density, is of special interest. Consider $g(t) = 0.5e^{-|t|}$, $t \in R$. For a parametric model $f_k(\mathbf{x}, \theta_k)$, based on Z^n, the maximum likelihood estimator of θ_k minimizes $\sum_{i=1}^{n} |Y_i - f_k(\mathbf{x}, \theta_k)|$ and σ is estimated by $\hat{\sigma} = (1/n)\sum_{i=1}^{n} |Y_i - f_k(\mathbf{x}, \hat{\theta}_{k,n})|$. 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.

3.3. A risk bound for ARM

Regarding the ARM algorithm, Yang (2001a) gave a risk bound. We improve the result by weakening the assumptions and, more importantly, with explicit constants in the risk bound.

Condition 1: There exists a constant $\tau > 0$ such that for all $i \ge 1$, with probability one, $\sup_{j \ge 1} \| \widehat{f}_{j,i} - f \|_\infty \le \sqrt{\tau}\sigma$.

Condition 2: There exist constants $0 < \xi_1 \le 1 \le \xi_2 < \infty$ such that $\xi_1 \le \widehat{\sigma}_{j,i}^2/\sigma^2 \le \xi_2$ with probability one for all $j \ge 1$ and $i \ge 1$.

The above conditions are satisfied if the regression function and the error variance are upper and lower bounded by known constants and the estimators are restricted accordingly. The boundness assumptions on the regression function and/or the error variance are commonly used in nonparametric regression (e.g., Juditsky and Nemirovski (2000)). Note that the constants τ, ξ_1 and ξ_2 are not used in the combining algorithm.

As in Yang (2001a), for the theoretical result, we study a slightly different estimator from those given earlier. For $i = n/2+1$, let $W_{j,i} = 1/K$ and for $n/2+1 < i \le n$, let

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

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

$$\widetilde{f}_n(\mathbf{x}) = \sum_{j=1}^{K} \widetilde{W}_j \widehat{f}_{j,n/2}(\mathbf{x}). \tag{1}$$

For simplicity, we only give the result with Gaussian errors here.

Theorem 1. *Assume that the errors are Gaussian and that Conditions* 1 *and* 2 *are satisfied. Then the risk of the combined regression estimator satisfies*

$$E \parallel \widetilde{f}_n - f \parallel^2 \le (1+\xi_2+9\tau/2) \inf_{j\ge 1} \Big(\frac{4\sigma^2 \log K}{n} + \frac{1}{\xi_1} E \parallel \widehat{f}_{j,n/2} - f \parallel^2 + \frac{C(\xi_1,\xi_2)}{\sigma^2} E(\widehat{\sigma}^2_{j,n/2} - \sigma^2)^2 \Big),$$

where $C(\xi_1,\xi_2) = (1/\xi_2 - 1 + \log \xi_2)/\xi_1^2 (1/\xi_2 - 1)^2$.

Remarks.

1. 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 holds regardless of whether there is a true model or not. 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 "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.
2. Theorem 1 deals with the total squared L_2 risk of the combined estimator. Bias properties of the estimator are also of interest, but are not directly addressed in this work.

Regarding the constant $C(\xi_1,\xi_2)$, for example, when $\xi_1 = 1/\xi_2 = 1/2$, $C(\xi_1,\xi_2) \approx 3.1$. From the result, up to a constant factor and an additive penalty $(\log K)/n$, the combined procedure achieves the best performance among $\widehat{f}_{j,n/2}$ plus the risk of variance estimation. Note that when ξ_1 and ξ_2 are around 1 and when τ is not large, the multiplicative factor is very reasonable. Roughly speaking if, when the sample size n increases, the estimators chosen to be combined are more and more accurate so that $\tau \to 0$ and ξ_1 and ξ_2 converge to 1, then basically the multiplicative factor is 2.

Theorem 1 improves the result for ARM in Yang (2001a) in two directions. First, with improved techniques, the constants in the performance bound are now explicitly given and sensible; second, Condition 1 is weaker than before.

Note that the estimator $\widetilde{f}_n$ at (1) is not the same as the $\widetilde{f}_n$ given in Section 3.1. The modified estimator is slightly more complicated and computationally more costly (but with the theoretical bound). As in Yang (2001a), the simpler one is recommended in practice.

3.4. Combining AIC and BIC as an illustration of ARM

As is well-known, neither AIC nor BIC performs better all the time. Roughly speaking, in terms of estimation accuracy, 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. Barron, Yang and Yu (1994) showed that, in theory, 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 in rates both for some parametric families and for some nonparametric classes. More recently, Yu and Hansen (1999) proposes a different MDL criterion to bridge AIC and BIC. They showed that the procedure is both consistent (as BIC) and asymptotically optimal (as AIC).

The algorithm ARM can be directly used to combine AIC and BIC in the hope that it will work well regardless of which one is better in terms of risks. Assume, for example, Gaussian errors and consider parametric families $f_k(\mathbf{x}, \theta_k)$, $k = 1, \ldots, K$. Let $\hat{\theta}_{k,i}$ and $\hat{\sigma}_{k,i}$ be the 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 as described in Section 3.2.

4. Selection Versus Combining

To our knowledge, there is little theoretical development in the literature on the difference between model/procedure selection and combining (mixing). General statistical risk bounds or asymptotic properties have been derived for both model selection (e.g., Shibata (1981), Li (1987), Barron and Cover (1991), Yang and Barron (1998), Barron, Birgé and Massart (1999), Lugosi and Nobel (1999) among many others) and model/procedure combining (Yang (2000ab, 2001a), Catoni (1999)), though model selection theories usually require the models to be finite-dimensional and is therefore more restrictive in some sense. These results typically imply that when the list of models/procedures is chosen appropriately,

both selection and combining result in function estimators that converge at optimal rates (or even with the right constant for some cases). For the purpose of understanding the advantage/disadvantage of selection versus combining, however, these results provide little insight.

There are two different directions in combining models/procedures. One is combining for adaptation (see Yang (2001a)), which intends to capture the best performance among the candidate models/procedures. The other is combining for improvement (see, Juditsky and Nemirovski (2000) and Yang (2002)), which intends to have a better performance than any of the original candidate model/procedures. For example, suppose that one regression procedure works well when the regression function is monotone and another procedure works well when the regression function is periodic. Then, if the true regression function happens to be decomposable into monotone and periodic components, combining the two procedures properly has a great potential to outperform the individual ones.

Intuitively, combining for improvement works when each model/procedure captures only part of the characteristics of the true model. It perhaps can be argued then that a better model/procedure should be constructed that reflects the characteristics in all the models/procedures (e.g., using mixture models). Indeed, some researchers hold this view and challenge the legitimacy of combining models (see, e.g., Clements and Hendry (1998, Chapter 10)). For the purpose of understanding the difference between selection and combining, this scenario addresses only an easy part of the matter: combining gives an opportunity to share different characteristics but selection does not. In other words, the advantage of combining here comes from better approximation capability. The challenging part of the matter is still unclear: assuming that one model/procedure performs the best among all the linear combinations (i.e., there is no approximation advantage in linear combining), how does (linear) combining compare to selection? For example, consider subset models or nested models in linear regression. Clearly, there is no gain in approximation capability when the models are to be combined. Is there any advantage for combining over selection here?

The problem is technically challenging. In this work, in a simple setting, we show that combining can be essentially better than selection. In fact, an estimator based on optimal testing will be shown to be worse than a combined estimator.

Obviously, selection is a special case of combining with combining weights concentrating on a single model/procedure. In this sense, one cannot show advantage of selection over combining. Nonetheless, one could study the advantage

of a selection method compared to a particular mixing strategy. Simulation study results in that direction will be given in the next section.

4.1. Inferiority of every estimator based on selection in a simple setting

In this subsection, we show in a simple setting that estimation based on selection performs worse compared to combining or mixing. In fact, estimation based on mixing with weights determined as in ARM will be shown to perform better.

For simplicity, consider a density estimation context. Let p_1 and p_2 be probability density functions on the real line. Let $X_1, \ldots, X_n$ be i.i.d. observations with the population density p. Suppose that it is known that p is either p_1 or p_2. We consider squared L_2 loss for the estimation of p.

More specifically, for convenience, let $p_1(x) = (1/\sqrt{2\pi\sigma^2})\exp(-(x-\theta_1)^2/2\sigma^2)$ and $p_2(x) = (1/\sqrt{2\pi\sigma^2})\exp(-(x-\theta_2)^2/2\sigma^2)$ where $\theta_2 > \theta_1$ and σ^2 are all given (known).

Note that, in this case, a selection rule between p_1 and p_2 corresponds to a testing rule of the hypothesis $H_0 : p = p_1$ versus $H_1 : p = p_2$. Let ϕ denote a simple testing rule and let C_ϕ and A_ϕ be its rejection and acceptance regions, respectively. Then the estimator based on testing (selection) is $\widehat{p}_\phi(x) = p_1(x)I_{A_\phi} + p_2(x)I_{C_\phi}$. The squared L_2 risk of this estimator is

$$R(p, \widehat{p}_\phi, n) = E_p \int (p(x) - \widehat{p}_\phi(x))^2 \, dx.$$

One particular testing rule is of interest. Let $L = \Pi_{i=1}^n p_2(X_i)/\Pi_{i=1}^n p_1(X_i)$ be the likelihood ratio statistic. Let ϕ^* be the rule that rejects H_0 when $L \geq 1$, or equivalently, when $\overline{X}_n \geq (\theta_1 + \theta_2)/2$. Clearly it is most powerful at its size.

Another approach to the estimation of p is by mixing p_1 and p_2 with a weight to p_1 of

$$W_1 = \frac{\exp\Big(-\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^n (X_i - \theta_1)^2\Big)}{\exp\Big(-\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^n (X_i - \theta_1)^2\Big) + \exp\Big(-\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^n (X_i - \theta_2)^2\Big)}.$$

The estimator is then defined by $\widetilde{p}_n(x) = W_1 p_1(x) + (1 - W_1)p_2(x)$. Its risk is $R(p, \widetilde{p}_n, n) = E_p \int (p(x) - \widetilde{p}_n(x))^2 \, dx$.

We have the following proposition that compares the two estimators based on selection and combining.

Proposition 1. *For any testing rule ϕ, we have* $\max_{i=1,2} R(p_i, \widehat{p}_\phi, n)/R(p_i, \widetilde{p}_n, n) > 1$ *for all n. In fact,* $\max_{i=1,2} R(p_i, \widehat{p}_\phi, n)/R(p_i, \widetilde{p}_n, n)$ *is minimized when $\phi = \phi^*$*

and then

$$\frac{R(p_1, \widehat{p}_{\phi^*}, n)}{R(p_1, \widetilde{p}_n, n)} = \frac{R(p_2, \widehat{p}_{\phi^*}, n)}{R(p_2, \widetilde{p}_n, n)} > 1.27 \text{ for all } n.$$

The proposition suggests the potential advantage of combining (mixing) over selection. For this simple case, combining with the weighting method as in ARM is superior to that based on the most powerful testing ϕ^*. We should also point out that there are additional issues that are worth consideration. If there are unknown parameters, a disadvantage may arise toward combining. The rough reasoning is that for selection, the whole data are used in the selection criterion (with parameters estimated by all the observations), but for assigning the weights for ARM, only half of the data are used in estimation (at the loss of some estimation accuracy) while the rest are used for performance assessment.

When multiple models are present with unknown parameters, the comparison between selection and combining becomes very difficult to analyze theoretically. Simulations in the next section show advantages in various scenarios.

4.2. Identifying the true model is not necessarily the right thing to do

Identifying the true model that governs the data, if possible, is an important task. When regression function estimation or prediction is the goal, the true model, even if assumed reasonably simple and known, may not perform the best. The well-known trade-off between bias and variance may prefer an incorrect but simpler model (see, e.g., Chapter 1 of Miller (1990) for an illustration in the context of prediction at a given site considering least squares estimators for the models). Along this line, we consider the global squared L_2 risk in a slightly different setting.

Example 1. Suppose the data come from the linear model

$$Y_i = \theta_0 + \theta_1\phi_1(X_i) + \cdots + \theta_k\phi_k(X_i) + \varepsilon_i,$$

where $\phi_0 = 1, \phi_1, \ldots, \phi_k$ are orthonormal with respect to the design distribution of X. Due to orthonormality, it is natural to consider the projection estimators of the linear coefficients:

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\phi_j(X_i) \qquad \text{for } 0 \le j \le k.$$

Since some of the true coefficients may be zero or small, there is potential advantage in considering subset models. Let Λ denote a subset of $\{0, \ldots, k\}$ and let $\widehat{f}_\Lambda(x) = \sum_{j\in\Lambda} \widehat{\theta}_j\phi_j(x)$. Then the risk is

$$R(f, \widehat{f}_\Lambda, n) = \sum_{j\in\Lambda} E\left(\widehat{\theta}_j - \theta_j\right)^2 + \sum_{j\notin\Lambda} \theta_j^2.$$

Let $M_j = E(\phi_j(X) - 1)^2$. By a simple calculation, $E\left(\widehat{\theta}_j - \theta_j\right)^2 = 1/n \sum_{l \neq j} \theta_l^2 + (\theta_j^2 M_j)/n + \sigma^2/n$. Suppose the true regression function f satisfies $f(x) = \sum_{j=0}^{m} \theta_j \phi_j(x)$ with non-zero θ_j's in the expression for some $0 < m \leq k$. Let $\Lambda^* = \{0, \ldots, m\}$. Consider a subset $\Lambda = \{0, \ldots, m-1\}$. Then

$$R(f, \widehat{f}_\Lambda, n) - R(f, \widehat{f}_{\Lambda^*}, n) = \theta_m^2 - 1/n \sum_{l \neq m} \theta_l^2 - \frac{\theta_m^2 M_m}{n} - \frac{\sigma^2}{n}.$$

Thus the wrong model corresponding to Λ can have a smaller risk compared to the use of the true model if the missing coefficient θ_m^2 is smaller than $1/n \sum_{l \neq m} \theta_l^2 + (\theta_m^2 M_m)/n + \sigma^2/n$. Note that this happens when $|\theta_m|$ is smaller than $\frac{\sigma}{\sqrt{n}}$ (n is small or σ is large), or when the sum of squares of the other coefficients is larger than θ_m^2.

From the example, for the purpose of estimating f or prediction, uncertainty in identifying the *true* model is not the issue per se. Even if the true model is known, it is not necessarily best to use it for prediction. Generally speaking, the best model (with the smallest risk) depends on the true coefficients, sample size, and noise level in a complex way.

5. Simulations

We consider several simulation settings to compare model selection with model combining, and to illustrate the advantage of ARM. The study was carried out using Splus.

5.1. Two non-nested models

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

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

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 (one based on selection and the other based on ARM) are simulated as the average of the squared differences between the true regression function and the estimates 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 parentheses in the tables are the corresponding standard errors.

Table 1. Comparing selection and combining for non-nested models.

	$\sigma^2 = 0.1$	0.3	0.5	1.0	1.5	2.0	3.0
Risk of Sel.	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-sel. 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)

From Table 1, 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. 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.

We mention that a plot (not included in this paper) of the risk ratio of selection versus combining (by ARM) against σ confirms what we have seen in Table 1: when σ is small, selection and combining perform very similarly; when σ increases, selection begins to perform better than combining but the advantage vanishes as σ increases further; then combining performs increasingly better than selection when the noise level gets higher.

5.2. Nested models

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

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

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 Section 3. As mentioned in the introduction, bagging unstable 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, up to 33% and 21% respectively. For AIC, however, bagging actually increases the risk (quite 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_{bag} better than AIC_{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_{bag} and BIC_{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, again demonstrating the usefulness of ARM for combining estimation procedures.

Table 2. Selection vs mixing when the true model has 4 terms.

	$\sigma^2 = 0.1$	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_{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_{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 3. Selection vs mixing when the true model has 2 terms.

	$\sigma^2 = 0.1$	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)

To show that the above cases are not atypical, we randomly generate the true model. Each of the five models receives 1/5 probability and the coefficients are independently generated with uniform distribution on $[-1, 1]$. Figure 1 gives the box-plot of the squared L_2 loss of the procedures (AIC, AIC_{bag}, BIC, BIC_{bag}, AIC-BIC combined and ARM) from 100 runs at each of six values of σ^2. The graph agrees with the tables well. Figure 2 compares the risks (based on 100 replications) of AIC, BIC and ARM with the true model randomly generated as

described above from 100 runs at four values of σ^2. The advantage of ARM is clearly seen when σ^2 is not small.

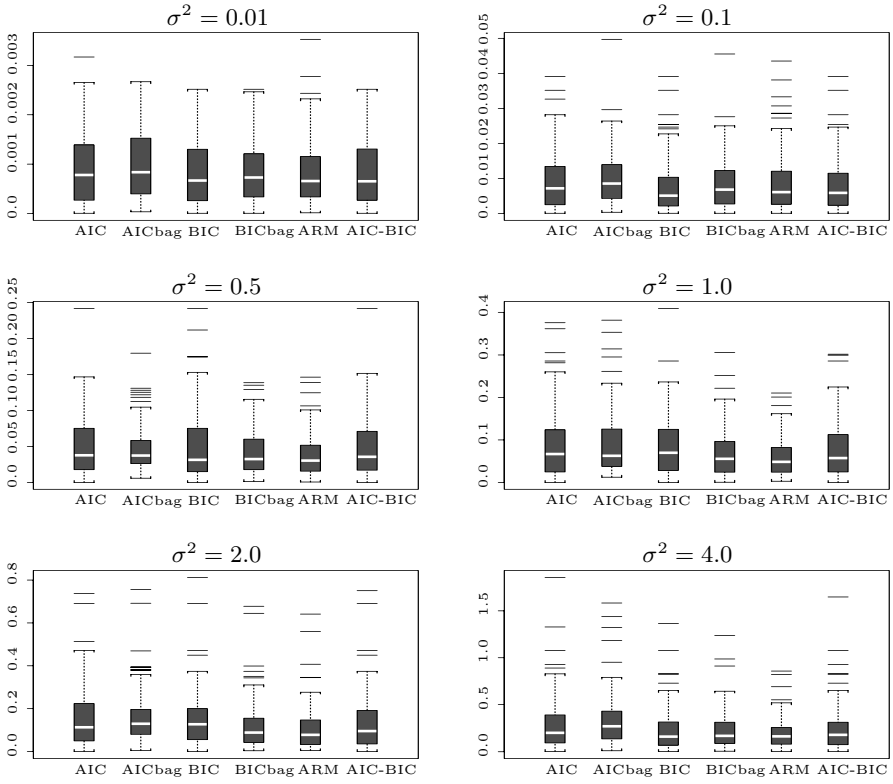

Figure 1. Comparing the squared L_2 losses of the procedures.

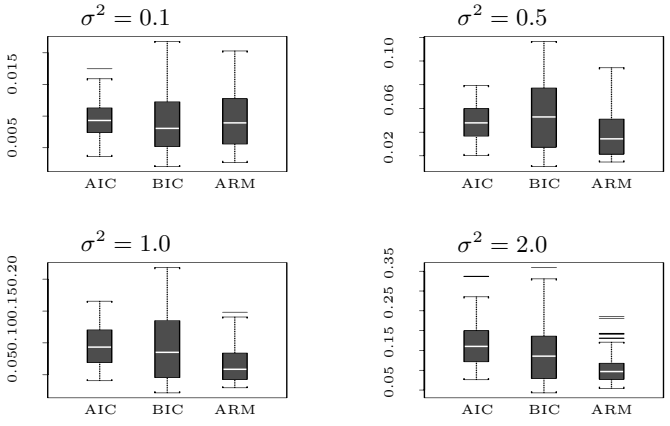

Figure 2. Comparing risks of AIC, BIC and ARM.

5.3. L_1-regression

Given in Table 4 is the result of a simulation to compare the performance of model selections and ARM under double-exponential errors. The first four models of the previous 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 $M = 200$.

Table 4. Selection vs mixing with double-exponential errors.

	$\sigma = 0.2$	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)

The advantage of ARM is again 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 tabled as

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

When σ^2 is 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).

5.4. Combining subset models when the number of predictors is small

Suppose there are four independent predictors uniformly distributed in [0,1]. Consider all subset models and use ARM and some BMA techniques to combine them. The true model 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.

Table 5. Comparing ARM with BMA.

	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)

In this simulation, ARM does substantially better than BMA for all four cases. The risk improvement is 4%, 22%, 24% and 29%, respectively.

In addition, we compare ARM with BMA in a random setting. We randomly choose one of the four models considered above with equal probability and then generate the coefficients (including the intercept) independently, all with the uniform distribution on $[-1, 1]$. Figure 3 gives the box-plots of the risks of BMA_{ow}, BMA and ARM based on 100 runs at three different noise levels.

From Figure 3, overall speaking, when σ equals 0.5 and 1, ARM performs substantially better than the BMA methods. When σ equals 1.5, the means of the risks of BMA and ARM are not significantly different. However, clearly, ARM continues to give more consistent risks than the BMA methods.

5.5. Crime data

Consider a crime data set 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 from 47 states in the US. The response variable is the crime rate and there are 15 candidate predictors. As in those works, log-transformation of Y and the predictors was applied and linear models were considered.

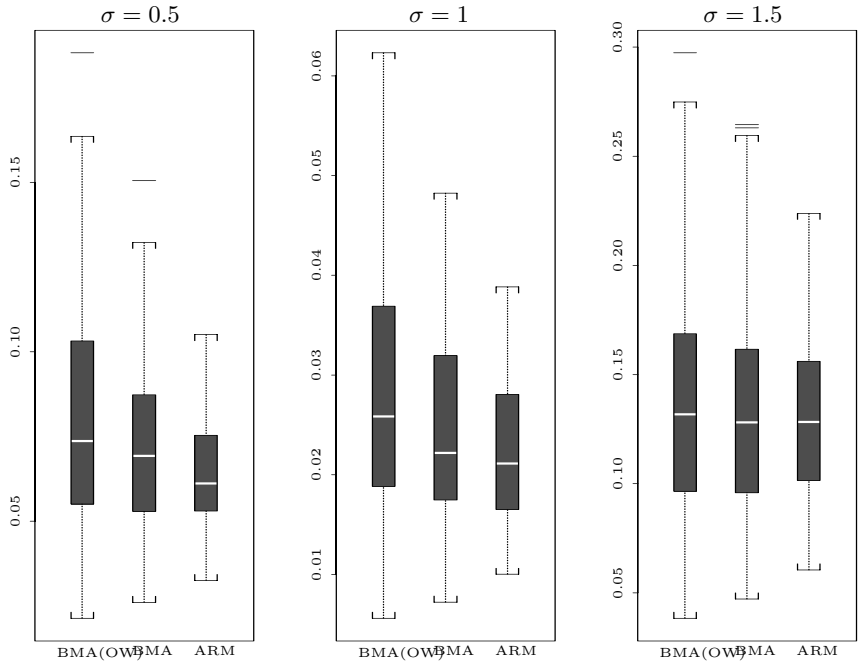

Figure 3. Comparing risks of BMA and ARM.

Raftery, Madigan and Hoeting (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 test 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 performing not as well as intended, 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 Splus), here we combine some plausible models to reduce computational cost. The stepwise (forward) selection method is used to order the predictors according to their order of appearance. Then the corresponding nested model is combined with ARM. Table 6 summarizes the PMSE's of the different methods based on 200 independent runs.

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%.

Table 6. Comparing BMA, Efroymson's method and ARM on a crime data.

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

6. Concluding Remarks

For estimating the regression function, approaches based on model/procedure selection and combining have better performance in different scenarios. Similarly to model selection theories (but less restrictive in some aspects), the risk bound for ARM shows that the estimators based on combining converges automatically at the best rate offered by the individual procedures.

Simulation results clearly suggest the 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 stabilizing the estimators based on model selection. When the error variance is smaller, so that there is not much difficulty comparing models, AIC and BIC can outperform ARM (for this case, bagging can also increase the risk significantly). In our experiments in this work, when the error variance is small, ARM and BIC behave equally well. The simulation also suggests 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. Simulations in this paper support this view.

For applications, one does not know beforehand if selection or mixing is better. It is tempting to construct an estimation procedure that automatically 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 would be helpful. 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 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. Simulation results showed the advantage of ARM over BMA methods based on BIC approximation in terms of estimation/prediction accuracy under squared error loss. To deal with a lot of candidate models, perhaps non-uniform initial weights on models could be incorporated in ARM to regulate the comparison.

We should also point out some disadvantages of ARM: it is difficult to interpret the estimate; the estimation is computer-intensive and more complex to program than AIC and BIC; and when the sample size is small, splitting the data may cause problems in estimation (e.g., having more parameters than the number of observations).

7. Proof of Theorem 1

Proof of Theorem 1. Let n_1 and n_2 be the sizes of the estimation and evaluation portions of the data, here $n_1 = n_2 = n/2$. Let $\widehat{f}_j$ denote $\widehat{f}_{j,n_1}$ and $\widehat{\sigma}_j^2$ denote $\widehat{\sigma}_{j,n_1}^2$ for $j \geq 1$. For simplicity in notation, in this proof, we drop the bold face format for a vector. Let

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

$$\begin{aligned} q^{n_2} &= \frac{1}{K}\sum_{j=1}^{K} \Pi_{i=n_1+1}^{n} \frac{1}{\sqrt{2\pi\widehat{\sigma}_j^2}} \exp\left(-\frac{1}{2\widehat{\sigma}_j^2}(y_i - \widehat{f}_j(x_i))^2\right) \\ &= \frac{1}{K}\sum_{j=1}^{K} \frac{1}{\left(2\pi\widehat{\sigma}_j^2\right)^{n_2}} \exp\left(-\frac{1}{2}\sum_{i=n_1+1}^{n} \frac{(y_i - \widehat{f}_j(x_i))^2}{\widehat{\sigma}_j^2}\right). \end{aligned}$$

Consider $\log\left(p^{n_2}/q^{n_2}\right)$. By monotonicity of the log function, for each fixed $j^* \geq 1$, we have

$$\log\left(p^{n_2}/q^{n_2}\right) \leq \log\left(\frac{\left(2\pi\sigma^2\right)^{-n_2/2} \exp\left(-\frac{1}{2}\sum_{i=n_1+1}^{n} \frac{(y_i - f(x_i))^2}{\sigma^2}\right)}{\frac{1}{K}\left(2\pi\widehat{\sigma}_{j^*}^2\right)^{-n_2/2} \exp\left(-\frac{1}{2}\sum_{i=n_1+1}^{n} \frac{(y_i - \widehat{f}_j(x_i))^2}{\widehat{\sigma}_{j^*}^2}\right)}\right)$$

$$= \log K + \frac{1}{2}\sum_{i=n_1+1}^{n}\left(\log\frac{\widehat{\sigma}_{j^*}^2}{\sigma^2} + \frac{(y_i-\widehat{f}_j(x_i))^2}{\widehat{\sigma}_{j^*}^2} - \frac{(y_i-f(x_i))^2}{\sigma^2}\right). \quad (2)$$

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

$$E_{n_1}\left(\log\frac{\widehat{\sigma}_{j^*}^2}{\sigma^2} + \frac{(y_i-\widehat{f}_j(x_i))^2}{\widehat{\sigma}_{j^*}^2} - \frac{(y_i-f(x_i))^2}{\sigma^2}\right) = \frac{\parallel \widehat{f}_j - f \parallel^2}{\widehat{\sigma}_{j^*}^2} + \frac{\sigma^2}{\widehat{\sigma}_{j^*}^2} - 1 - \log\frac{\sigma^2}{\widehat{\sigma}_{j^*}^2}. \quad (3)$$

On the other hand, observe that q^{n_2} is equal to

$$\begin{aligned}
&\frac{1}{K}\sum_{j=1}^{K}\frac{1}{\sqrt{2\pi\widehat{\sigma}_j^2}}\exp\left(-\frac{1}{2\widehat{\sigma}_j^2}(y_{n_1+1}-\widehat{f}_j(x_{n_1+1}))^2\right)\\
&\times\frac{\displaystyle\sum_{j=1}^{K}\frac{1}{\sqrt{4\pi^2\widehat{\sigma}_j^4}}\exp\left(-\frac{1}{2\widehat{\sigma}_j^2}(y_{n_1+1}-\widehat{f}_j(x_{n_1+1}))^2-\frac{1}{2\widehat{\sigma}_j^2}(y_{n_1+2}-\widehat{f}_j(x_{n_1+2}))^2\right)}{\displaystyle\sum_{j=1}^{K}\frac{1}{\sqrt{2\pi\widehat{\sigma}_j^2}}\exp\left(-\frac{1}{2\widehat{\sigma}_j^2}(y_{n_1+1}-\widehat{f}_j(x_{n_1+1}))^2\right)}\\
&\times\cdots\times\frac{\displaystyle\sum_{j=1}^{K}\frac{1}{\Pi_{i=1}^{n}\sqrt{2\pi\widehat{\sigma}_j^2}}\exp\left(-\sum_{i=n_1+1}^{n}\frac{1}{2\widehat{\sigma}_j^2}(y_i-\widehat{f}_j(x_i))^2\right)}{\displaystyle\sum_{j=1}^{K}\frac{1}{\Pi_{i=1}^{n-1}\sqrt{2\pi\widehat{\sigma}_j^2}}\exp\left(-\sum_{i=n_1+1}^{n-1}\frac{1}{2\widehat{\sigma}_j^2}(y_i-\widehat{f}_j(x_i))^2\right)}.
\end{aligned}$$

Let $p_i = (1/\sqrt{2\pi\sigma^2})\times\exp(-(y_i-f(x_i))^2/2\sigma^2)$ and $g_i = \sum_{j=1}^{K} W_{j,i}(1/\sqrt{2\pi\widehat{\sigma}_j^2})\times\exp(-(y_i-\widehat{f}_j(x_i))^2/2\widehat{\sigma}_j^2)$ for $n_1+1\le i\le n$. It follows by the definition of $W_{j,i}$ that $\log(p^{n_2}/q^{n_2}) = \sum_{i=n_1+1}^{n}\log\left(\frac{p_i}{g_i}\right)$. Together with (2) and (3), under the i.i.d. assumption on the data, we have

$$\sum_{i=n_1+1}^{n} E\log\left(\frac{p_i}{g_i}\right) \le \log K + \frac{n_2}{2}E\left(\frac{\parallel \widehat{f}_j - f \parallel^2}{\widehat{\sigma}_{j^*}^2} + \frac{\sigma^2}{\widehat{\sigma}_{j^*}^2} - 1 - \log\frac{\sigma^2}{\widehat{\sigma}_{j^*}^2}\right). \quad (4)$$

Now observe that, conditioned on the first part of the data and x_i as denoted by $E_{n_1}^{'}$ below, we have

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

where the inequality is the familiar relationship between the Kullback-Leibler divergence and the squared Hellinger distance. The Hellinger distance is lower

bounded in terms of the difference of their means as follows (see Lemma 1 of Yang 2001a). Let p and g be two probability densities on the real line with respect to a measure ν, with means μ_p and μ_g, variances $0 < \sigma_p^2 < \infty$ and $0 < \sigma_g^2 < \infty$ respectively. Then

$$\int (\sqrt{p} - \sqrt{g})^2 \, d\nu \geq \frac{(\mu_p - \mu_g)^2}{2\left(\sigma_p^2 + \sigma_g^2\right) + (\mu_p - \mu_g)^2}.$$

Under Conditions 1 and 2, it is straightforward to verify that the variance of g_i is upper bounded by $\xi_2\sigma^2 + 4\tau\sigma^2$. Since the mean of g_i (as a density function in y_i) is $\widehat{s}_i(x_i) = \sum_{j=1}^K W_{j,i}\widehat{f}_j(x_i)$, we have

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

Together with (4),

$$\begin{aligned}&\sum_{i=n_1+1}^{n} E\Big(\frac{(\widehat{s}_i(X_i) - f(X_i))^2}{\sigma^2\,(2(1+\xi_2) + 9\tau)}\Big)\\ &\leq \log K + \frac{n_2}{2} E\Big(\frac{\parallel \widehat{f}_j - f \parallel^2}{\widehat{\sigma}_{j^*}^2} + \frac{\sigma^2}{\widehat{\sigma}_{j^*}^2} - 1 - \log\frac{\sigma^2}{\widehat{\sigma}_{j^*}^2}\Big).\end{aligned}$$

By convexity, we have

$$E\Big(\Big(\frac{1}{n_2}\sum_{i=n_1+1}^{n} \widehat{s}_i(X_i)\Big) - f(X_i)\Big)^2 \leq \frac{1}{n_2}\sum_{i=n_1+1}^{n} E\,(\widehat{s}_i(X_i) - f(X_i))^2.$$

Note that $\frac{1}{n_2}\sum_{i=n_1+1}^{n} \widehat{s}_i(x) = \widetilde{f}_n(x)$. Thus

$$\begin{aligned}&E \parallel \widetilde{f}_n - f \parallel^2\\ &\leq \sigma^2\,(2(1+\xi_2) + 9\tau)\,\Big(\frac{\log K}{n_2} + \frac{1}{2}E\Big(\frac{\parallel \widehat{f}_j - f \parallel^2}{\widehat{\sigma}_{j^*}^2} + \frac{\sigma^2}{\widehat{\sigma}_{j^*}^2} - 1 - \log\frac{\sigma^2}{\widehat{\sigma}_{j^*}^2}\Big)\Big).\end{aligned}$$

It is straightforward to verify that if $x \geq x_0 > 0$, $x - 1 - \log x \leq c_{x_0}(x-1)^2$ for a constant $c_{x_0} = (x_0 - 1 - \log x_0)/\,(x_0 - 1)^2$. Together with the fact that the above inequality holds for every j^*, under Condition 2, it follows that

$$\begin{aligned}&E \parallel \widetilde{f}_n - f \parallel^2\\ &\leq (1 + \xi_2 + 9\tau/2)\inf_{j\geq 1}\left(\frac{4\sigma^2\log K}{n} + \frac{1}{\xi_1}E \parallel \widehat{f}_j - f \parallel^2 + \frac{C(\xi_1,\xi_2)}{\sigma^2}E(\widehat{\sigma}_{j^*}^2 - \sigma^2)^2\right),\end{aligned}$$

where $C(\xi_1,\xi_2) = (1/\xi_2 - 1 + \log\xi_2)/\xi_1^2\,(1/\xi_2 - 1)^2$. The conclusion follows.

Proof of Proposition 1. By symmetry, $R(p_1, \widetilde{p}_n, n) = R(p_2, \widetilde{p}_n, n)$, and given by

$$E_{\theta_1} \int (p_1(x) - \widetilde{p}_n(x))^2 \, dx = E_{\theta_1} (1 - W_1)^2 \cdot \int (p_1(x) - p_2(x))^2 \, dx.$$

For the estimator based on selection, the risks are closely related to the probabilities of the two types of errors. Indeed, we have

$$R(p_1, \widehat{p}_\phi, n) = E_{\theta_1} \int (p_1(x) - \widehat{p}_\phi(x))^2 \, dx = P_{\theta_1}(C_\phi) \int (p_1(x) - p_2(x))^2 \, dx,$$
$$R(p_2, \widehat{p}_\phi, n) = E_{\theta_2} \int (p_2(x) - \widehat{p}_\phi(x))^2 \, dx = P_{\theta_2}(A_\phi) \int (p_1(x) - p_2(x))^2 \, dx.$$

Let C_{ϕ^*} denote the rejection region $\{\overline{X}_n \geq (\theta_1 + \theta_2)/2\}$ of ϕ^*. The probability of type I error is $\alpha_n = P_{\theta_1}(\overline{X}_n \geq (\theta_1 + \theta_2)/2) = 1 - \Phi(\sqrt{n}(\theta_2 - \theta_1)/2\sigma)$. By the Neyman-Pearson Lemma, this is the most powerful test of size α_n. Due to symmetry of ϕ^*, the probability of type II error is also α_n. It follows that for any test ϕ of size less than α_n, the probability of type II error is necessarily larger than α_n. As a consequence, the maximum of probabilities of Type I and II error is minimized when ϕ is ϕ^*. Then the first statement in Proposition 1 follows from the second statement.

It remains to show $R(p_1, \widehat{p}_\phi, n)/R(p_1, \widetilde{p}_n, n) > 1.27$ for all n. Note that

$$\begin{aligned}
E_{\theta_1}(1 - W_1)^2 &= E_{\theta_1} \left(\frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta_2)^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta_1)^2\right) + \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta_2)^2\right)} \right)^2 \\
&= E_{\theta_1} \left(\frac{1}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \theta_2)^2\right) + 1} \right)^2 \\
&= E_{\theta_1} \left(\exp\left(\frac{(\theta_1 - \theta_2)\, n\overline{X}_n}{\sigma^2} + \frac{n(\theta_2^2 - \theta_1^2))}{2\sigma^2} \right) + 1 \right)^{-2} \\
&= E \left(\exp\left(\frac{\sqrt{n}(\theta_2 - \theta_1)}{\sigma} \left(\frac{\sqrt{n}(\theta_2 - \theta_1)}{2\sigma} - Z \right) \right) + 1 \right)^{-2},
\end{aligned}$$

where Z in the last equality denotes a random variable with a standard normal distribution. Let $\beta = \sqrt{n}(\theta_2 - \theta_1)/\sigma$. Then the risk ratio of selection versus combining is

$$\frac{R(p_1, \widehat{p}_\phi, n)}{R(p_1, \widetilde{p}_n, n)} = \frac{1 - \Phi\left(\frac{\beta}{2}\right)}{E\left(\exp\left(\beta\left(\frac{\beta}{2} - Z\right)\right) + 1\right)^{-2}}.$$

The expression seems very complicated to evaluate analytically. Numerical calculations show that the ratio is decreasing in $\beta > 0$ (with derivative approaching zero) and is lower bounded by 1.27 and upper bounded by 2 (asymptotically achieved when $\beta \to 0$, i.e., when σ is large relative to the sample size and the difference of the means). This completes the proof of Proposition 1.

Acknowledgements

The author thanks the reviewers and an associate editor for their comments, which improved the presentation of the paper. This research was partially supported by National Security Agency Grant MDA9049910060 and National Science Foundation CAREER Grant DMS0094323.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Internat. Symp. Inform. Theory* (Edited by B. N. Petrov, F. Csaki and Akademia Kiado), 267-281. Budapest.

Barron, A. R. (1987). Are Bayes rules consistent in information? In *Open Problems in Communication and Computation* (Edited by T. M. Cover and B. Gopinath), 85-91. Springer-Verlag, New York.

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

Barron, A. R., Yang, Y. and Yu, B. (1994). Asymptotically optimal function estimation by minimum complexity criteria. In *Proc.* 1994 *Internat. Symp. Inform. Theory*, 38. Trondheim, Norway.

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

Breiman, L. (1996a). Stacked regressions. *Mach. Learning* **24**, 49-64.

Breiman, L. (1996b). Bagging predictors. *Mach. Learning* **24**, 123-140.

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

Catoni, O. (1999). "Universal" aggregation rules with exact bias bounds. Preprint.

Cesa-Bianchi, N., Freund, Y., Haussler, D. P., Schapire, R. and Warmuth, M. K. (1997). How to use expert advise? *J. Assoc. Comput. Mach.* **44**, 427-485.

Clements, M. P. and Hendry, D. F. (1998). *Forecasting economic time series*. Cambridge University Press, New York.

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

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer, New York.

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

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

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

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*, p.63. Santa Fe, NM.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* (with discussions) **14**, 382-417.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.* **28**, 681-712.

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

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

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

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Inform. and Comput.* **108**, 212-261.

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

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

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

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

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92**, 179-191.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statist. Math.* **35**, 415-423.

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

Yang, Y. (2000a). Mixing strategies for density estimation. *Ann. Statist.* **28**, 75-87.

Yang, Y. (2000b). Combining different regression procedures for adaptive regression. *J. Multivariate Anal.* **74**, 135-161.

Yang, Y. (2001a). Adaptive regression by mixing. *J. Amer. Statist. Assoc.* **96**, 574-588.

Yang, Y. (2001b). Combining forecasting procedures: some theoretical results. Accepted by *Econom. Theory*.

Yang, Y. (2002). Aggregating regression procedures for a better performance. Revised for *Bernoulli*.

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

Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.

E-mail: yyang@iastate.edu

(Received September 2001; accepted February 2003)