Adaptive Regression by Mixing

Article *in* Journal of the American Statistical Association · February 2001

DOI: 10.2307/2670298 · Source: RePEc

1 author:

Yuhong Yang
Tsinghua University
127 PUBLICATIONS **5,452** CITATIONS

Adaptive Regression by Mixing *

Yuhong Yang
Department of Statistics
Iowa State University
Ames, IA 50011-1210, USA
yyang@iastate.edu

Abstract

Adaptation over different procedures is of practical importance. Different procedures perform well under different conditions. In many practical situations, it is rather hard to assess which conditions are (approximately) satisfied so as to identify the best procedure for the data at hand. Thus automatic adaptation over various scenarios is desirable.

A practically feasible method, named Adaptive Regression by Mixing (ARM) is proposed to convexly combine general candidate regression procedures. Under mild conditions, the resulting estimator is theoretically shown to perform optimally in rates of convergence without knowing which of the original procedures work the best.

Simulations are conducted in several settings, including comparing a parametric model with nonparametric alternatives, comparing a neural network with a projection pursuit in multi-dimensional regression, and combining bandwidths in kernel regression. The results clearly support the theoretical property of ARM.

The ARM algorithm assigns weights on the candidate models/procedures via proper assessment of performance of the estimators. The data are split into two parts, one for estimation and the other for measuring behavior in prediction. While there are many plausible ways to assign the weights, ARM has a connection with information theory, which ensures the desired adaptation capability. Indeed, under mild conditions, we show that the squared L_2 risk of the estimator based on ARM is basically bounded above by the risk of each candidate procedure plus a small penalty term of order $1/n$. Minimizing over the procedures gives the automatically optimal rate of convergence for ARM.

Model selection often induces unnecessarily large variability in estimation. Alternatively, a proper weighting of the candidate models can be more stable, resulting in a smaller risk. Simulations suggest that ARM works better than model selection using AIC or BIC when the error variance is not very small.

Keywords: Adaptive estimation, combining procedures, nonparametric regression.

1 Introduction

In this work, we propose a method, ARM, to combine estimators of a regression function based on the same data. The focus is on adaptation with respect to the estimators. ARM is intended to be a theoretically proven and practical algorithm to combine regression procedures.

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

Regression estimation includes parametric and nonparametric approaches. Parametric modelings have advantages of simplicity, better interpretability of the regression function, and high efficiency when the chosen models are appropriate. On the other hand, with a few parameters, a parametric model may not provide enough flexibility needed to describe the underlying regression function well. As a solution, nonparametric methods have been introduced. Nonparametric approaches include smoothing (e.g., kernel, smoothing spline, local polynomial) and parametric approximation (e.g., in terms of polynomials, trigonometrical, wavelets, polynomial splines and others). For the case of high dimensional regression, to overcome the curse of dimensionality, various dimension reduction methods (e.g., additive models, neural nets, regression trees, projection pursuit) have been proposed.

While various methods are available, in applications, one faces the problem of choosing the right method for the data at hand. The problem includes: 1). comparison among parametric models; 2). comparison between a parametric model and a nonparametric procedure; 3). comparison among nonparametric procedures.

For the first case above, in addition to the traditional hypothesis testing techniques to compare nested models, a variety of model selection criteria have been proposed and used in practice, including AIC (Akaike (1973)), BIC (Schwartz (1978)), and cross-validation (Stone (1974)). Recent general model selection theories focusing on nonparametric estimation include Barron and Cover (1991) using a minimum description length criterion, Yang and Barron (1998) using a penalized likelihood criterion, Yang (1999a) using a penalized least squares criterion with model complexity incorporated to handle a large number of candidate models, Barron, Birgé and Massart (1999) with a unified theoretical treatment on penalization methods, Lugosi and Nobel (1999) using an empirical complexity penalized criterion, and others. These results basically show that for nonparametric estimation, with an appropriate model selection criterion, the estimator performs as well (often in terms of rate of convergence) as if one were told before hand which model is the best (in statistical risk) for the data. For comparing a parametric model with a nonparametric alternative, lack-of-fit tests based on nonparametric smoothing have been studied (see, e.g., Hart (1997)). To our knowledge, no formal general methods have been proposed to compare different nonparametric procedures.

Despite the derived nice theoretical properties, model selection often produces a rather unstable estimator in applications. A small perturbation of the data can result in selecting a very different model. As a consequence, estimators of the regression function based on model selection often have a rather unnecessarily large variance. Breiman (1996b) proposed a resampling method named *bagging* to stabilize procedures and reported dramatic improvement in simulations and some data examples.

When estimating the regression function is the goal, an alternative to model selection is model averaging. Intuitively, when two models are really close in terms of a selection criterion, appropriate weighting of the models can be much better than an exaggerated 0-1 decision ("winner takes all" in some sense). Bayesian model averaging methods have been proposed based on Bayesian considerations with a recent focus on the case when a large number of models are to be combined. For references and interesting results in that direction, the readers are referred to a review article by Hoeting, Madigan, Raftery and Volinsky (1999). The goal in this paper is to provide a practically feasible weighting method with a proven theoretical property in terms of statistical risk. It will be applicable for combining parametric and/or nonparametric procedures without the restriction to model-based estimates.

Ideas of combining procedures based on the same data have been considered in several scientific research fields. In statistics, a few methods have been proposed to linearly combine function estimators. Olkin and Spiegelman (1987) suggest convexly combining a parametric and a kernel estimates for density estimation. Recent methods include cross-validation based "stacking" by Wolpert (1992) and Breiman (1996a), a bootstrap based method by LeBlanc and Tibshirani (1996), a stochastic approximation based method by Juditsky and Nemirovski (2000), and information-theoretic based methods to combine density (or conditional probability) estimators by Yang (2000ac) and Catoni (1997). As pointed out by LeBlanc and Tibshirani, the idea of stacking was considered earlier by Stone (1974) in the name of "model-mix". Juditsky and Nemirovski proposed algorithms and derived interesting theoretical upper and lower bounds for linear aggregation. Yang and Catoni both show that in the context of density estimation, with proper weighting, the combined procedure has a risk bounded above by a multiple of the smallest risk over the original procedures plus a small penalty, which usually ensures the optimal rate of convergence for estimating the unknown density function. In another direction, mixtures-of-experts (Jacobs, Jordan, Nowlan, and Hinton (1991)) and hierarchical mixtures-of-experts models (Jordan and Jacobs (1994)) provide a way to convexly combine certain parametric models (e.g., generalized linear models) with flexible localized weights. The approximation and estimation properties were studied by Jiang and Tanner (1999, 2000). In the field of forecasting, various researches demonstrate usefulness of combining forecasts, see, e.g., Clemen (1989) for a review of work in that direction. A randomizing method was proposed by Foster and Vohra (1993) to combine forecasts and it was shown that the combined forecast does asymptotically as well as the best individual forecast in a certain probability sense. In information theory, universal coding has been studied for many years in the spirit of adaptation. See Merhav and Feder (1998) and Barron, Rissanen and Yu (1998) for reviews of related work in that field. In recent years, combining procedures becomes a very active topic in computational learning theory. The focus

has been on deriving mixed strategies with optimal performance in terms of a cumulative loss without any probabilistic assumptions at all on the generation of the data. Algorithms have been proposed and the combined strategies are shown to have cumulative loss upper bounded by that of the best individual procedure (losing a constant factor sometimes) plus a penalty of order of the logarithm of the number of procedures being combined, see, e.g., Vovk (1990, 1998), Littlestone and Warmuth (1994), Cesa-Bianchi *et al* (1997), Haussler, Kivinen and Warmuth (1998), Kivinen and Warmuth (1999), and Cesa-Bianchi and Lugosi (1999).

The present paper is built upon an earlier work of the author (Yang (2000b)), where it is shown that given a collection of regression procedures, under Gaussian errors, a suitably combined procedure usually behaves asymptotically as well as the best procedure in terms of rate of convergence. Though pointing at applications, the theoretical algorithm used to combine procedures there is unfortunately not feasible for implementation. In this work, we propose a practical algorithm with theoretically proven properties. It does not require normality and it can be used when there are multiple candidate error distributions. Simulations suggest its great potential in applications.

The paper is organized as follows. In Section 2, we present the algorithm ARM. Its generalization to handle multiple candidate error distributions is given in Section 3. Illustrations with simulation and real data are provided in Section 4. General theoretical developments on ARM are presented in Section 5. An additional theoretical result for homoscedastic errors is given in Section 6. A conclusion follows in Section 7. The proofs of the main results are in Section 8.

2 Algorithm ARM

Consider the regression setting

$$Y_i = f(X_i) + \sigma(X_i) \cdot \varepsilon_i,\ i = 1, ...n,$$

where $(X_i, Y_i)_{i=1}^n$ are i.i.d. copies from the joint distribution of (X, Y) with $Y = f(X) + \sigma(X) \cdot \varepsilon$. The explanatory variable X could be multidimensional and has an unknown distribution P_X. The error component ε is assumed to be independent of X and has a known probability density $h(t)$, $t \in R$ (with respect to a measure μ) with mean 0 and a finite variance. The unknown function $\sigma(x)$ controls the variance of the random error given $X = x$. Our goal is to estimate the regression function f based on the data $Z^n = (X_i, Y_i)_{i=1}^n$.

Suppose a finite collection of regression procedures have been proposed for estimating f. Here a regression procedure (or strategy), say, δ, refers to a method of estimating f and σ based on Z^n at each

sample size n. Let δ_j, $1 \le j \le J$ denote the proposed regression procedures. Let $\hat{f}_{j,i}(x) = \hat{f}_{j,i}(x; Z^i)$ and $\hat{\sigma}_{j,i}(x) = \hat{\sigma}_{j,i}(x; Z^i)$ $(i \ge 1)$ denote the estimator of f and σ respectively by procedure δ_j based on Z^i . No special requirement will be put on the procedures and they could be proposed under completely different assumptions on the regression function (e.g., smoothness, monotonicity, additivity). An example for multi-dimensional regression is that δ_1 is a simple linear regression, δ_2 is an additive modeling, δ_3 is a projection pursuit, and δ_4 is a neural network method with the number of nodes selected by AIC. Some of the procedures could be of the same type but with different choices of hyper-parameters (e.g., cubic splines versus quadratic splines). The procedures could share variance estimators if desired.

We propose the following algorithm, which we call Adaptive Regression by Mixing (ARM), for combining the procedures. For simplicity in notation, assume n is even.

Algorithm ARM

Step 0. Randomly permute the order of the observations.

Step 1. Split the data into two parts $Z^{(1)} = (X_i, Y_i)_{i=1}^{n/2}$ and $Z^{(2)} = (X_i, Y_i)_{i=n/2+1}^{n}$.

Step 2. Obtain estimates $\hat{f}_{j,n/2}(x; Z^{(1)})$ of f based on $Z^{(1)}$ for $1 \le j \le J$. Estimate the variance function $\sigma^2(x)$ by $\hat{\sigma}^2_{j,n/2}(x)$.

Step 3. For each j, evaluate predictions. For $n/2 + 1 \le i \le n$, predict Y_i by $\hat{f}_{j,n/2}(X_i)$. Compute

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

Step 4. Compute the current weight for procedure δ_j. Let

$$W_j = \frac{E_j}{\sum_{l=1}^{J} E_l}.$$

Step 5. Repeat steps 0-4 $(M-1)$ more times and average the weights over the M random permutations. Let $\hat{W}_j$ denote the weight of procedure δ_j obtained this way. The final estimator is

$$\hat{f}_n(x) = \sum_{j=1}^{J} \hat{W}_j \hat{f}_{j,n}(x).$$

Note that the first half of the data is used for estimation by each procedure and the second half of the data is used to assess its prediction performance and then the weights are assigned accordingly. While there are many plausible ways for choosing the weights (e.g., using a model selection criterion as in Buckland *et al* (1996)), the weighting method of ARM has a connection with information theory, which ensures a strong theoretical property as will be shown in Section 5. The random permutations (with M suitably large) average out the variability in data splitting.

Two choices of h, namely, Gaussian and the double-exponential densities are of special interest. For the Gaussian case, the quantity E_j in Step 3 becomes

$$E_j = \frac{(2\pi)^{-n/4} \exp\left(- \sum_{i=n/2+1}^{n} \frac{(Y_i - \hat{f}_{j,n/2}(X_i))^2}{2\hat{\sigma}^2_{j,n/2}(X_i)} \right)}{\Pi_{i=n/2+1}^{n} \hat{\sigma}_{j,n/2}(X_i)}.$$

If the variance estimator is a constant function (which is suitable when $\sigma(x)$ is a constant), then E_j reduces to $(2\pi)^{-n/4}\hat{\sigma}_{j,n/2}^{-n/2} \exp\left(- \left(\sum_{i=n/2+1}^{n} (Y_i - \hat{f}_{j,n/2}(X_i))^2 \right) / \left(2\hat{\sigma}^2_{j,n/2} \right) \right)$. The numerator of the exponent is exactly the sum of squares in prediction. If the procedures are maximum likelihood estimation based on parametric families of the regression function, the estimators are the familiar least squares estimators. The double-exponential density is

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

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

The ARM algorithm can be computationally intensive when M is large since each procedure is applied M times. As will be seen from the simulations in Section 4, the computational cost may well be worthwhile to obtain adaptivity in estimation.

3 ARM with multiple candidate error distributions

The ARM procedure in the previous section requires knowledge of the error distribution up to a scaling parameter. For many applications, one may not be certain about the error distribution and therefore might want to try different specifications of the error distributions. For example, Gaussian, t and double-exponential distributions can be considered to provide flexibility in handling error distributions with different degrees of heavy-tailedness.

Let $\mathcal{H} = \{h_1, ..., h_I\}$ be a collection of candidate error distributions. Suppose that for each $1 \le j \le J$, δ_j is a regression procedure associated with one choice, say h_{k_j} with $k_j \in \{1, ..., I\}$. For this case, to combine the estimators, one just needs to modify the weight assignment on the procedures by using the corresponding error distribution, i.e.,

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

As will be seen in Section 5, theoretically speaking, as long as the true error distribution is in $\mathcal{H}$, and it is used by at least one good procedure, then the final estimator will perform well adaptively.

For some applications, the candidate families of the regression function are of parametric forms, say $f_j(x;\theta_j)$ with unknown parameter θ_j. Suppose a few error distributions are plausible (e.g., Gaussian and double-exponential). Then a choice of a family $f_j(x;\theta_j)$ and an error distribution together determine a regression procedure based on the maximum likelihood estimation. These procedures can be combined as described above. A simulation study will be given in the next Section in this context, where ARM is seen to put very small weights on the procedures with the wrong error distribution and performs as well as if the true error distribution were known.

4 Experiments

In this section, we demonstrate applications of ARM both in simulations and on real data. Some simulation studies for comparing ARM with familiar model selection criteria AIC and BIC for parametric regressions are in Yang (1999b). Here to have a focused illustration, unless stated otherwise, we assume the errors are normally distributed with unknown variance σ^2 and the number of permutations for ARM is chosen to be 200. The squared L_2 loss is used as a measure of discrepancy in estimating the regression function. It is simulated using 1000 (unless stated otherwise) new independent draws from the distribution of X. All the simulations are conducted using Splus.

1. Parametric or nonparametric?

i). A simulation. For the first illustration, we consider three regression methods: a parametric modeling (quadratic regression), regression spline, and a nonparametric smoothing method (smoothing spline). For regression spline, we use cubic B-splines with the number of equally-spaced knots automatically selected. In some sense, the regression spline can be viewed as an intermediate method between parametric and fully nonparametric procedures. The true underlying regression function is one of the following functions on [0,1]:

$$\begin{aligned}
&\text{Case 1: } f(x) = (x-0.2)^2\\
&\text{Case 2: } f(x) = e^{-0.5+2x}\\
&\text{Case 3: } f(x) = e^{-0.5+3x}\\
&\text{Case 4: } f(x) = 1+2x^2+0.8e^{-200(x-0.5)^2}.
\end{aligned}$$

The sample size is taken to be 150 and $\sigma^2 = 1.0$ for the first three cases and $\sigma^2 = 0.5$ for Case 4. A typical realization of data for each case is plotted in Figure 1.

We use the least squares method for quadratic regression. The number of knots for the polynomial spline is selected by AIC. The smoothing spline estimation is provided in Splus with the function

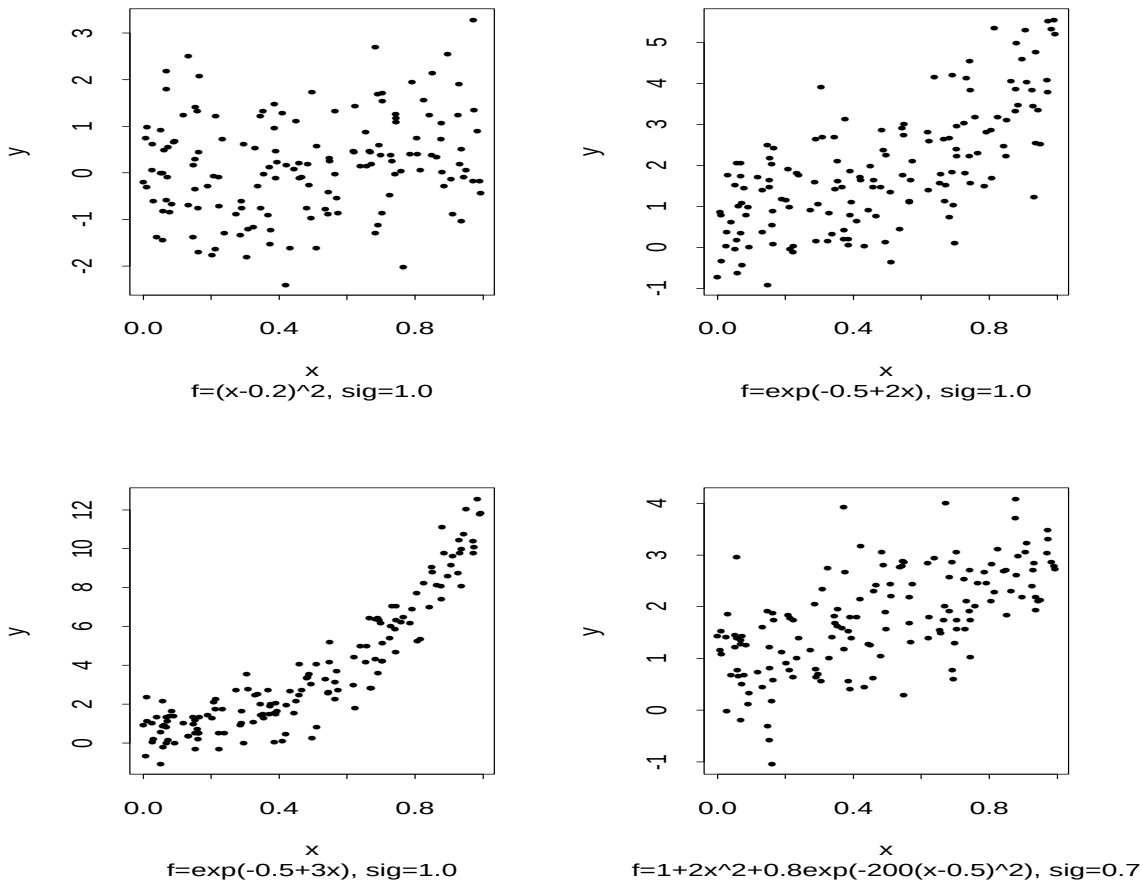

Figure 1. *Typical Realization of Data for Simulation 1*

`Smooth.Spline` and we use the default choice of cross-validation for choosing the smoothing parameter. We combine the three methods using ARM. The squared L_2 risks of the estimators are computed based on 100 replications. The numbers in the parentheses in Table 1 are the corresponding standard errors.

Quadratic regression works much better than the nonparametric alternatives for the first two cases, but becomes much worse for the latter two due to lack of flexibility. Note that even though the quadratic model is wrong for the second case, it still outperforms the nonparametric procedures. It is clear that the ARM estimator here behaves quite well adaptively for the four cases, as if it knew whether the parametric model or a nonparametric procedure is better.

ii). Real data. We consider two data sets: "Old Faithful Geyser" in Yellowstone National Park (222 observations) as studied in e.g., Simonoff (1996, Chapter 5), and "Babinet Point" data (355 observations) studied by Cleveland (1993) and Hart (1997). The data are plotted in Figure 2.

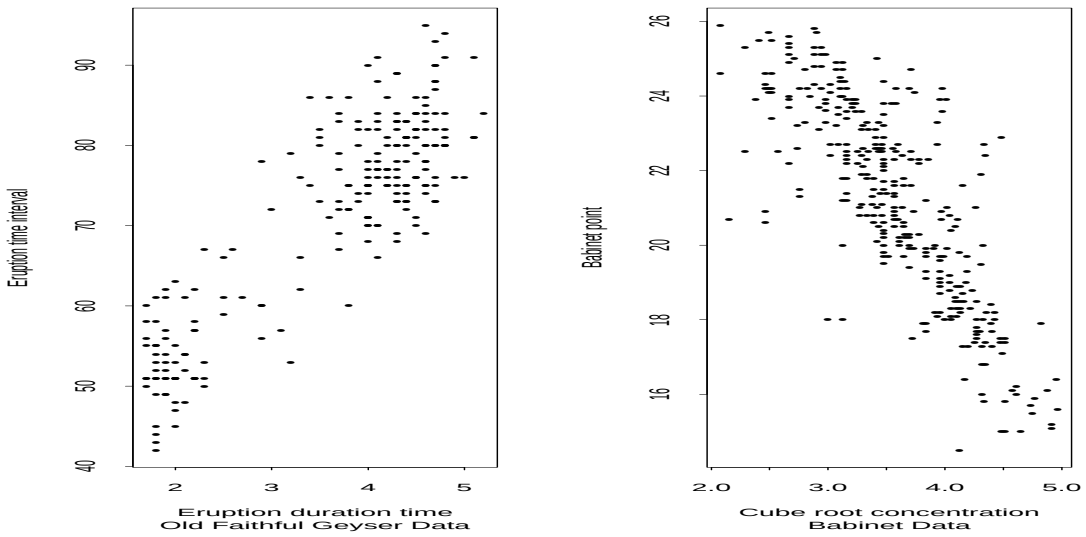

Figure 2. *Two Real Data Sets*

For both cases, there is an apparent linear trend. Nonparametric procedures have been considered alternatively for improvement. A lack-of-fit test in Hart (1997, Chapter 10) leads to rejection of the simple linear model with a very small p-value. For the purpose of estimating the regression function, one could consider combining a linear model and a nonparametric procedure, which is chosen to be a smoothing spline here. The variance parameter σ^2 is estimated by $RSS/(n - df)$ for smoothing spline, where RSS is the residual sum of squares and df is the degrees of freedom of the smoothing spline fit provided in the function `Smooth.Spline`.

For comparing the procedures, we randomly split the data into an estimation set (90%) and a test set (10%) and the average squared error in prediction is computed using the test data. Three hundred replications are conducted and the results are in Table 2.

	Case 1	Case 2	Case 3	Case 4
Quadratic	0.0195	0.0206	0.0938	0.0402
	(0.0012)	(0.0016)	(0.0019)	(0.0005)
PolySpline	0.0369	0.0351	0.0551	0.0225
	(0.0035)	(0.0035)	(0.0034)	(0.0009)
SmoothSpline	0.0283	0.0323	0.0450	0.0203
	(0.0027)	(0.0030)	(0.0039)	(0.0010)
ARM	0.0205	0.0231	0.0436	0.0202
	(0.0015)	(0.0017)	(0.0022)	(0.0009)

Table 1: Quadratic or Nonparametric?

	Geyser	Babinet
Linear	38.61 (0.52)	3.203 (0.041)
SmoothSpline	47.86 (0.92)	3.226 (0.047)
ARM	37.59 (0.52)	2.994 (0.038)

Table 2: Comparing Linear Model with Smoothing Spline for Two Data Sets

In our experiment, for the Babinet data, we found smoothing spline occasionally gives a very poor estimate, probably due to very uneven locations of the explanatory variable due to random splitting. For a fair comparison, we trim 5% of the replications when calculating the risks. We also tried using test sets consisting of every tenth observation in terms of X (ordered) at different start, and the outcomes are similar.

For the Geyser data, the simple linear model does much better than the nonparametric estimator (the prediction error is only 81% of the smoothing spline). For the Babinet data, the parametric and nonparametric procedures behave very similarly, even though the linear model is rejected with very strong evidence in Hart (1997). This seems to suggest that though closely related, testing and prediction can be very different objectives. The important point here is that for both cases, ARM performs as well as the better procedure. For both cases, the linear model gets larger weight with average (over the replications) 0.98 and 0.73 respectively.

2. Which bandwidth is right? For many nonparametric procedures, there are tuning parameters, which usually highly influence the performance. For example, bandwidth choice for kernel regression is crucial to have the right amount of smoothness needed for data. Here we give a simple demonstration using ARM to combine different bandwidths.

We consider 4 underlying regression functions on [0,1]: $f_1(x) = \exp(-x)$, $f_2(x) = 1+2x^2+\exp(-5(x-0.5)^2)$, $f_3(x) = 1+2x^2+\exp(-200(x-0.5)^2)$, and $f_4(x) = \exp(-200(x-0.2)^2)/\sqrt{0.005\pi}+\exp(-200(x-0.8)^2)/\sqrt{0.005\pi}$. The functions differ in smoothness. The sample size is set to be 200 and $\sigma^2 = 0.5$. Typical realizations of data are given in Figure 3.

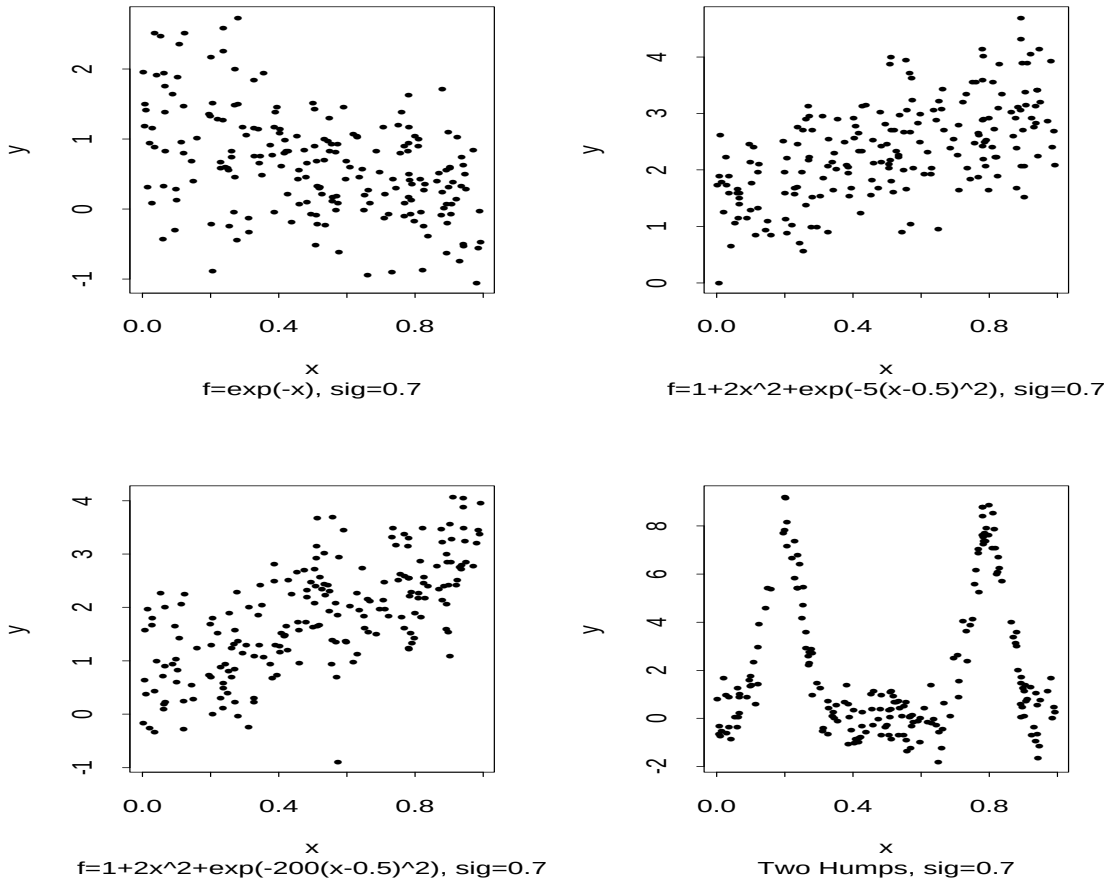

Figure 3. *Typical Realization of Data for Simulation 2*

We use the normal kernel for kernel regression and consider 6 different bandwidths: $h = 0.01$, 0.05, 0.1, 0.3, 0.5, 0.7. ARM is used to combine the corresponding kernel estimators. We use the same variance estimator $\hat{\sigma}^2 = 1/(2(n-1))\sum_{i=1}^{n-1}(Y_{(i+1)} - Y_{(i)})^2$ in ARM, where $Y_{(i)}$ denotes the response at the ith smallest value of X. The estimator $\hat{\sigma}^2$ converges at the parametric rate under minor smoothness conditions on f as will be mentioned in Section 6. The risks in Table 3 are based on 100 replications.

The average weights that ARM puts on the bandwidths are given as in Table 4 (the numbers in the parentheses are the corresponding standard deviations).

As expected, the best bandwidths are different for the four cases. From the above risk table, the risk of ARM is reasonably close to the best kernel estimator. Not knowing the best bandwidth, one expects to pay a price for adaptation. It is of interest in the future to study the influence of discretization accuracy on performance of ARM and compare it with kernel estimators based on different bandwidth selectors.

	$h = 0.01$	$h = 0.05$	$h = 0.1$	$h = 0.3$	$h = 0.5$	$h = 0.7$	ARM
Case 1	0.2121	0.0389	0.0199	0.0079	0.0064	0.0078	0.0072
	(0.0116)	(0.0039)	(0.0027)	(0.0017)	(0.0016)	(0.0018)	(0.0017)
Case 2	0.2129	0.0392	0.0196	0.0109	0.0241	0.0535	0.0139
	(0.0120)	(0.0034)	(0.0023)	(0.0022)	(0.0036)	(0.0050)	(0.0026)
Case 3	0.2134	0.0428	0.0252	0.0442	0.0719	0.1023	0.0294
	(0.0113)	(0.0041)	(0.0032)	(0.0028)	(0.0039)	(0.0055)	(0.0036)
Case 4	0.2427	0.1196	0.4738	4.2316	6.2888	7.0257	0.1210
	(0.0128)	(0.0119)	(0.0324)	(0.0489)	(0.0277)	(0.0204)	(0.0122)

Table 3: Combine Bandwidths with ARM

	$h = 0.01$	$h = 0.05$	$h = 0.1$	$h = 0.3$	$h = 0.5$	$h = 0.7$
Case 1	0.0015	0.0156	0.0744	0.2617	0.3166	0.3301
	(0.0045)	(0.0208)	(0.0557)	(0.0578)	(0.0376)	(0.0708)
Case 2	0.0015	0.0299	0.1833	0.4726	0.2410	0.0717
	(0.0027)	(0.0272)	(0.0802)	(0.0707)	(0.0538)	(0.0266)
Case 3	0.0016	0.0736	0.4732	0.3240	0.0897	0.0379
	(0.0021)	(0.0540)	(0.1507)	(0.1189)	(0.0630)	(0.0276)
Case 4	0.0359	0.9467	0.0130	0.0015	0.0015	0.0015
	(0.0364)	(0.0376)	(0.0153)	(0.0032)	(0.0032)	(0.0032)

Table 4: Average Weights on the Bandwidths by ARM

Sometimes bandwidth selectors based on classical model selection criteria such as AIC and cross-validation differ dramatically from plug-in methods (see, e.g., Loader (1999)). For such cases, as far as the estimation (or prediction) accuracy is concerned, one can combine the corresponding estimators using the ARM algorithm. It remains to be seen how well this works for applications.

3. Multi-dimensional regression. When the dimension d of the explanatory variable X gets high, regression estimation becomes much more difficult and dimension reduction regression procedures become very useful. As an illustration, we here consider $d = 4$ and two dimension reduction methods, namely, neural networks and projection pursuit. The simulation is conducted using the `ppreg` function in Splus and the Splus code `nnet` by Venable and Ripley (1997).

The input random variables $X_1, ..., X_4$ are independent and uniformly distributed in $[0, 1]$. The underlying regression function is chosen to be

$$f(x_1, x_2, x_3, x_4) = \frac{2\exp(x_1 + 0.5x_2 - 0.8x_3 + 0.1x_4)}{1 + \exp(x_1 + 0.5x_2 - 0.8x_3 + 0.1x_4)} - \frac{3\exp(0.5x_1 + 2x_2 + 0.1x_3 + 0.1x_4)}{1 + \exp(0.5x_1 + 2x_2 + 0.1x_3 + 0.1x_4)}.$$

The sample size is taken to be 400. The squared L_2 losses are simulated using 5000 new X values, and the risks are calculated with 100 replications. For the neural network, the option of size is chosen to be 3, and for the projection pursuit, the minimum and maximum number of terms are set to be 1

and 3 respectively. We consider small ($\sigma = 0.1$) and large ($\sigma = 1$) variances. For combining the two procedures, we use a three-stage ARM algorithm, which will be given in Section 6 using half the data for estimating f, a quarter of the data for estimating σ^2, and the remaining quarter for performance assessment for final weight assignment.

	$\sigma = 0.1$	$\sigma = 1$
NN	0.00065	0.11914
	(0.00002)	(0.00340)
PP	0.00141	0.04990
	(0.00002)	(0.00287)
ARM	0.00065	0.06604
	(0.00002)	(0.00244)

Table 5: Comparing Neural Networks with Projection Pursuit

It is interesting to observe in Table 5 that when $\sigma = 0.1$, neural networks behave much better, but the projection pursuit becomes much better when $\sigma = 1$. In both cases, ARM again automatically behaves close to the better one as intended.

4. Normal or double-exponential error? Here we demonstrate that ARM with multiple candidate error distributions can be adaptive with respect to the error distribution as well. For simplicity, consider 6 independent and uniformly distributed random input variables $X_1, ..., X_6$ on [0,1] and consider the six nested linear models with X_1 only, X_1 and X_2, ..., and $X_1, ..., X_6$. The true model is:

$$Y = X_1 + 0.9X_2 + 0.7X_3 + \sigma\varepsilon,$$

where ε has a double exponential distribution. The sample size is 50, and 100 replications are used to simulate the risks. Let AIC_N and AIC_{DE} denote AIC based on the Gaussian error and double-exponential error respectively. Similarly denote BIC and ARM combining the 6 families under a choice of the error distribution. We use ARM_{Both} to denote the ARM estimator with both the normal and double-exponential errors considered as in Section 4 (each combination of a family and a choice of error distribution gives a model).

From Table 6, when the error distribution is misspecified, AIC, BIC, and ARM behave substantially worse respectively. ARM with both normal and double-exponential errors considered perform almost identically as if only the double-exponential (the correct error) were considered. In fact, the models based on normal error receive extremely small weights. When σ is small, ARM and BIC performs similarly, and ARM becomes advantageous when σ gets larger with risk reduction close to 40% compared to AIC and BIC when $\sigma = 1$. We have also tried true models with different number of terms and the results are

	AIC_N	AIC_{DE}	BIC_N	BIC_{DE}	ARM_N	ARM_{DE}	ARM_{Both}
$\sigma = 0.1$	0.0021	0.0017	0.0018	0.0013	0.0019	0.0012	0.0012
	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0001)	(0.0001)
$\sigma = 0.5$	0.0601	0.0404	0.0531	0.0349	0.0477	0.0310	0.0310
	(0.0057)	(0.0038)	(0.0045)	(0.0033)	(0.0038)	(0.0023)	(0.0023)
$\sigma = 1$	0.1901	0.1566	0.2174	0.1538	0.1213	0.0938	0.0938
	(0.0167)	(0.0136)	(0.0149)	(0.0126)	(0.0085)	(0.0064)	(0.0064)

Table 6: Combine Models Considering Both Normal and Double-Exponential Errors

typically similar. When σ is not small, even if the number of terms in the true model is extreme (1 or 6) (which seems to favor model selection since there is no chance to underfit or overfit respectively), ARM still does a better or similar job compared to AIC and BIC.

5. Discussion. Based on the simulation studies above, it seems that ARM indeed provides adaptation capability over various models/procedures as intended. We focused on the case when a small number of models/procedures are to be combined. It is also of interest to study the performance of ARM when there is a large number of candidate procedures such as in the case of subset selection and in some cases of discretization of continuous tuning parameters (e.g., multidimensional bandwidth). In addition, for ARM here, the data are plit evenly for estimation and predication for convenience. A natural question is: What is the optimal splitting of the sample size? One does not expect to have a universally good choice, but data-driven methods may well lead to performance improvement over the half-half splitting. See Section 5 for related discussion on this issue from a theoretical point of view. Another issue worth future investigation is the choice of M. For the simulations, focusing on the main issue of adaptation, we choose M to be rather large ($M = 200$). The standard errors of the ARM estimators compared to the original procedures suggest that such a choice enables ARM to be very stable. To reduce the computation cost, a suitable criterion may be used to stop permutation in ARM when the weights for the procedures become stabilized in a certain measure.

Model selection and hypothesis testing have been applied widely in statistical applications. They are very useful and important tools in identifying a good model, which allows proper interpretation of the characteristics of the regression function important to the subject of study. Model averaging, however, is not appropriate for that task. When estimating f or future prediction is of primary interest, minimizing a statistical risk rather than a choice of a model becomes the appropriate goal. Generally speaking, when the noise level is very low relative to the signal, there is little difficulty identifying the best model and accordingly model selection does a very good job in terms of prediction as well. When the noise level becomes higher, models are harder to be distinguished and the discrete decision of selecting one

becomes inferior to model averaging with appropriate weighting, as shown in Section 4.4 above. Yang (1999b) gives more demonstrations with parametric models, where it is also seen that ARM still performs better even if AIC and BIC are stabilized using the bagging procedure proposed by Breiman (1996b) (In fact, in our simulations with nested models, bagging sometimes actually degraded the performance dramatically.) For hypothesis testing, e.g., testing a parametric family of f against a nonparametric family, the outcome of rejecting H_0 does not necessarily mean that the parametric estimator performs worse than a nonparametric alternative in estimation or prediction. Even if a parametric model is known to be wrong, it can still result in a better risk than a nonparametric procedure, depending on the degree of departure from the parametric family and the complexity of the nonparametric procedure relative to the parametric estimator. Our simulations suggest, paying the price of more intensive computing and increased difficulty in interpretation, when it is hard to compare the candidate procedures, ARM does a better or much better job in estimating the regression function than approaches based on model selection or hypothesis testing.

Finally we briefly discuss the difference between *combining for adaptation* (as in our work) and *combining for improving the individual procedures* (e.g., as studied in Juditsky and Nemirovski (2000)). The former intends to automatically capture the best performance among the procedures being considered, and the latter intends to improve the individual procedures through a (generally nontrivial) linear combination. While being more aggressive with the potential of beating the best original procedure, as one might expect, the latter needs to pay a higher price compared to ARM. Indeed, as will be shown in Section 5, the estimator based on ARM usually automatically converges at the best rate offered by the procedures, but as shown in Juditsky and Nemirovski (2000), targeting at the best linear combination of the original estimators, in general, any combined estimator can not be uniformly within a smaller order than $1/\sqrt{n}$ in squared L_2 distance from the best linear combination.

5 A general theory for ARM

In this section, we present a theoretical result on ARM in a general form. The result basically shows that ARM indeed provides adaptivity among procedures as intended.

Let $\| f - g \| = \left(\int |f(x) - g(x)|^2 dP_X \right)^{1/2}$ be the L_2 distance between two functions f and g with respect to the distribution of X. We consider the squared loss $\| f - \hat{f} \|^2$ as a global measure of performance for the theoretical development.

Let $\Delta = \{\delta_j, j \geq 1\}$ be a collection of regression procedures with δ_j producing an estimator $\hat{f}_{j,i}$ based on Z^i for each $i \geq 1$. Here we allow a countable collection of procedures to be combined in

theory. The index set $\{j \geq 1\}$ is allowed to degenerate to a finite set (as is considered in the previous sections). The risk of a procedure δ for estimating f at sample size n is denoted $R(f;n;\delta)$, i.e., $R(f;n;\delta) = E \parallel f - \hat{f}_n \parallel^2$ with the expectation taken under the regression function f and the variance function σ^2.

We have the following assumptions for our results. We first assume h is fixed (known) with mean 0 and variance 1.

A1. The regression function $f(x)$ is uniformly bounded, i.e., $\parallel f \parallel_\infty \leq A < \infty$, and $\sigma(x)$ is uniformly bounded above and below, i.e., $0 < \underline{\sigma} \leq \sigma(x) \leq \overline{\sigma} < \infty$ for some constants $A, \underline{\sigma}$, and $\overline{\sigma}$. We assume the estimators produced by δ_j's also satisfy these requirements respectively.

A2. The error distribution h is such that for each pair $0 < s_0 < 1$ and $T > 0$, there exists a constant B (depending on s_0 and T) such that

$$\int h(x) \log \frac{h(x)}{\frac{1}{s} h\left(\frac{x-t}{s}\right)} \mu(dx) \leq B\left((1-s)^2 + t^2\right)$$

for all $s_0 \leq s \leq s_0^{-1}$ and $-T < t < T$.

For Assumption A1, the constants A, $\underline{\sigma}$, $\overline{\sigma}$ are involved in the derivation of the risk bounds, but they need not to be known to perform the ARM procedures. The Assumption A2 is mild and is satisfied by Gaussian, t (with degrees of freedom bigger than 2), double-exponential, and many other distributions.

The following construction is similar in spirit to that given in Section 2, but is more general.

For each n, choose an integer $N = N_n$ with $1 \leq N_n \leq n$. Unless stated otherwise, N_n is chosen to be of order n. Let

$$W_{j,n-N+1} = \pi_j, \quad j = 1, 2, ...,$$

where π_j's are positive numbers summing up to one, i.e., $\sum_{j=1}^{\infty} \pi_j = 1$. They can be viewed as initial weights (or prior probabilities) of the procedures in Δ (the role of π_j's will be discussed later after Theorem 1). For $n - N + 2 \leq i \leq n$, let

$$W_{j,i} = \frac{\pi_j \Pi_{l=n-N+1}^{i-1} h\left(\frac{Y_{l+1} - \hat{f}_{j,l}(X_{l+1})}{\hat{\sigma}_{j,l}(X_{l+1})}\right) / \hat{\sigma}_{j,l}(X_{l+1})}{\sum_{k=1}^{\infty} \pi_k \Pi_{l=n-N+1}^{i-1} h\left(\frac{Y_{l+1} - \hat{f}_{k,l}(X_{l+1})}{\hat{\sigma}_{k,l}(X_{l+1})}\right) / \hat{\sigma}_{k,l}(X_{l+1})}. \tag{1}$$

Note that $\sum_{j \geq 1} W_{j,i} = 1$ for each $i = n - N + 1, ..., n$. Let

$$\widetilde{f}_i(x) = \sum_j W_{j,i} \hat{f}_{j,i}(x)$$

and define

$$\overline{\hat{f}}_n(x) = \frac{1}{N} \sum_{i=n-N+1}^{n} \widetilde{f}_i(x). \tag{2}$$

Note that $\overline{\hat{f}}_n(x)$ depends on the order of observations. Under the i.i.d. assumption on the data, the estimator can be improved by taking its conditional expectation given the values of observations (ignoring the order, as mentioned in Section 2). For applications, one can randomly permute the order a number of times and average $\overline{\hat{f}}_n(x)$ over the permutations to average out the order effect. For the theoretical development, we focus on $\overline{\hat{f}}_n$ as defined above. The risk bounds certainly apply to the improved estimator using permutations.

Let δ^* denote the procedure producing $\{\overline{\hat{f}}_n,\ n \geq 1\}$. We have the following performance bound on this general ARM.

Theorem 1: Assume that Conditions A1 and A2 are satisfied. For any given countable collection of estimation procedures $\Delta = \{\delta_j, j \geq 1\}$, we can construct a single combined estimation procedure δ^* as given above such that

$$R(f; n; \delta^*) \leq C_1 \inf_j \left(\frac{1}{N_n} \log \frac{1}{\pi_j} + \frac{C_2}{N_n} \sum_{l=n-N_n+1}^{n} \left(E\|\sigma^2 - \hat{\sigma}^2_{j,l}\|^2 + E\|f - \hat{f}_{j,l}\|^2 \right) \right), \tag{3}$$

where the constant C_1 depends on A and $\overline{\sigma}$, and C_2 depends on $A, \overline{\sigma}/\underline{\sigma}$ and h. This upper bound also applies to the average risk of $\widetilde{f}_i$, $n - N_n + 1 \leq i \leq n$, i.e.,

$$\frac{1}{N_n} \sum_{i=n-N_n+1}^{n} E\|f - \widetilde{f}_i\|^2 \leq C_1 \inf_j \left(\frac{1}{N_n} \log \frac{1}{\pi_j} + \frac{C_2}{N_n} \sum_{l=n-N_n+1}^{n} \left(E\|\sigma^2 - \hat{\sigma}^2_{j,l}\|^2 + E\|f - \hat{f}_{j,l}\|^2 \right) \right).$$

Remarks:

1. The risk bound is still valid if Δ and the weights π_j are chosen to depend on the sample size n.

2. As the risk bound suggests, variance estimation is also important for ARM. Even if a procedure estimates f very well, as seen from the definition of $W_{j,i}$ in (1), a bad estimator of σ^2 can substantially reduce its weight in the final estimator. See Section 6 for discussions on estimating the variance (parameter) in case of homoscedastic errors.

3. Note that the weights for combining the procedures in ARM are global (the weights do not depend on x). Mixtures-of-experts (Jacobs, Jordan, Nowlan, and Hinton (1991)) and hierarchical mixtures-of-experts models (Jordan and Jacobs (1994)) use localized weights to combine parametric models with flexibility and are shown to have nice theoretical properties in terms of approximation, convergence rate and asymptotic normality by Jiang and Tanner (1999, 2000).

4. Bounds of the type

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{y}^*_i)^2 \leq \frac{C \log M}{n} + C' \inf_{1 \leq j \leq M} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{y}_{j,i})^2$$

for combining M procedures in computational learning theory with related algorithms are given in e.g., Haussler, Kivinen and Warmuth (1998), Vovk (1998), and Kivinen and Warmuth (1999). Here $\widehat{y}_{j,i}$ is

the predicted value of Y_i by the j-th procedure based on previous observations and X_i, and $\widehat{y}_i^*$ is the combined forecast of Y_i. When $C' = 1$, taking expectation, the loss bound can yield risk bounds similar to those given in Theorem 1 (even with $C_1 C_2 = 1$) without the variance estimation component. However, the loss bound is obtained under the very strong assumption that Y_i's are uniformly bounded in a known range, which is usually not appropriate to model uncertainty in Y in statistics. When $C' > 1$, by taking expectation, the loss bound is not strong enough to derive a risk bound comparable to Theorem 1.

5. The ARM algorithm has a close relationship with data compression. Two facts are used in the derivation of the adaptation algorithm and the risk bound. The first is a connection between data compression and density estimation based on an i.i.d. sample, namely, if one can compress the data well one can estimate the underlying distribution (density) well and vise versa. For the task of data compression of i.i.d. observations coming from an unknown member in a given class of probability distributions, one is to find a joint distribution on the observations so that it is simultaneously close (under the Kullback-Leibler divergence) to all the product densities with the marginal density in the given class (see, e.g., Yang and Barron (1999, Section 3)). This establishes a correspondence between an estimation procedure and a joint distribution on the observations. The second fact is that averaging over multiple joint densities on the product space provides adaptation capability in data compression. That is, to have a data compression method suitable for multiple classes of distributions, one can simply take an average (with suitable weights if necessary) of the joint distributions on the product space which are constructed for each of the classes individually. Based on the two facts, to obtain adaptation with respect to multiple estimation procedures, one just need to average the corresponding joint distributions on the product space and then use the first fact in the other direction to obtain an adaptive estimator of the underlying distribution (density). For regression, some additional difficulties are involved compared to the case of density estimation.

6. For the estimators $\tilde{f}_i$, the risk bound given in the theorem is valid for the average of their risks. We do not have a risk bound individually at each sample size. However, by taking average of them as defined in (2), by convexity, one gets the risk bound in (3) for the new estimator $\overline{f}_n$ at sample size n. It is not clear to us how it compares with the simpler estimator $\tilde{f}_n$ in performance.

In Sections 2-4, it is assumed that the number of procedures to be combined, J, is finite and the equal weight $\pi_j = 1/J$ is used. In practice, when J is large, we may assign smaller weights for more complex estimation procedures (see, e.g., Yang and Barron (1998) for some natural assignments based on coding). Then the risk bound in (3) is a trade-off between accuracy and complexity. For a complex procedure (with a small prior weight), its role in the risk bound becomes significant only when the sample size

becomes large. It is of future interest to study how sensitive the prior weights are on the performance of ARM.

The risks of a good procedure for estimating f and σ^2 usually decrease as the sample size increases. For such a case, the influence of N_n on the quantity in the parenthesis inside (3) is clear: larger N_n decreases the penalty term $\frac{1}{N_n}\log\frac{1}{\pi_j}$ (for not knowing which procedure is the best) but increases the main terms involving the risks of the procedure. For familiar parametric and nonparametric estimators, the risks $E\|f-\hat{f}_{j,n}\|^2$ usually decrease around a polynomial order $n^{-r}\eta(n)$ for some $0<r\le 1$ and $\eta(n)$ (e.g., $\log n$) being a slowly changing function (variance estimation is similar). Then $(1/N_n)\sum_{l=n-N_n+1}^{n} E\|f-\hat{f}_{j,l}\|^2$ is of the same order as $n^{-r}\eta(n)$ for any choice of $N_n\le \tau n$ for some $0<\tau<1$ (the choice of $N_n=n$ results in an extra logarithmic factor for a parametric rate with $r=1$). If this holds for the good procedures in Δ, the risk bound in (3) becomes

$$R\left(f;n;\delta^*\right)\le C\left\{\inf_j\left(\frac{1}{N_n}\log\frac{1}{\pi_j}+E\|\sigma^2-\hat{\sigma}^2_{j,n}\|^2+E\|f-\hat{f}_{j,n}\|^2\right)\right\}.$$

A reasonable choice for N_n is $n/2$, for which case the penalty term $\frac{1}{N_n}\log\frac{1}{\pi_j}$ is of order $1/n$ and it does not affect the rate of convergence for regression estimation.

Let us briefly discuss an implication of the above risk bounds on adaptive rate of convergence. The risk bounds involve estimators of the variance function $\sigma^2(x)$. For variance estimation based on local polynomials, see Ruppert *et al* (1997). Under very minor smoothness condition on f, for homoscedastic errors (i.e., $\sigma^2(x)$ is a constant), estimators of σ^2 converging at rate $1/n$ in squared risk can be obtained independent of the models (see, Rice (1984)). Then the effect of variance estimation in the risk bound is of no larger order than the risk for estimating f by the procedures. Model dependent variance estimators can also be used so that the risk of variance estimation does not affect the rate of convergence for estimating f (see Section 6). Then $R\left(f;n;\delta^*\right)$ is bounded asymptotically by order $\inf_j\left(\frac{1}{N_n}\log\frac{1}{\pi_j}+E\|f-\hat{f}_{j,n}\|^2\right)$. A consequence is that when N_n is chosen to be of order n, the estimator $\hat{f}_n$ based on ARM converges automatically at the best rate of convergence offered by any of the procedures being considered. Some results on adaptive rate of convergence over function classes are in Yang (2000b).

Now let us relate Theorem 1 to the ARM algorithm given in Section 2. For each $n/2+1\le m\le n$, define

$$E_{j,m}=\frac{\Pi_{i=n/2+1}^{m} h\left(\frac{(Y_i-\hat{f}_{j,n/2}(X_i))}{\hat{\sigma}_{j,n/2}(X_i)}\right)}{\Pi_{i=n/2+1}^{m}\hat{\sigma}_{j,n/2}(X_i)}$$

and

$$W_{j,m}=\frac{E_{j,m}}{\sum_{l=1}^{J}E_{l,m}}.$$

Let

$$\widetilde{f}_m(x) = \sum_{j=1}^{J} W_{j,m} \hat{f}_{j,n/2}(x).$$

Note that $W_{j,m}$ here is the special case of (1) with $N = n/2$ and the use of $\hat{f}_{j,n/2}$ instead of $\hat{f}_{j,i}$ for $n/2+1 \le i \le n$ (i.e., without updating the estimators for saving the computational cost). Then we have the following result on the average risk for $\widetilde{f}_m$, $n/2+1 \le m \le n$.

Corollary 1: Under the same conditions as in Theorem 1, we have

$$\frac{1}{n/2} \sum_{m=n/2+1}^{n} E\|f - \widetilde{f}_m\|^2 \le C_1 \inf_j \left(\frac{1}{n/2} \log \frac{1}{\pi_j} + C_2 E\|\sigma^2 - \hat{\sigma}^2_{j,n/2}\|^2 + C_2 E\|f - \hat{f}_{j,n/2}\|^2 \right).$$

Proof: Take $\widetilde{f}_{j,m}(x; Z^m) = \hat{f}_{j,n/2}(x; Z^{n/2})$ and $\widetilde{\sigma}^2_{j,m}(x; Z^m) = \hat{\sigma}^2_{j,n/2}(x; Z^{n/2})$ for $n/2+1 \le m \le n$ (i.e., do not update the estimators for m in the specified range $[n/2+1, n]$). The risk bound then follows directly from Theorem 1. This completes the proof of Corollary 1.

If one uses $\widetilde{f}_n(x) = \sum_{l=1}^{J} W_{j,n} \hat{f}_{j,n}(x)$ instead of $\sum_{l=1}^{J} W_{j,n} \hat{f}_{j,n/2}(x)$, then the average risk is upper bounded by

$$\begin{aligned} & \frac{1}{n/2} \sum_{m=n/2+1}^{n} E\|f - \widetilde{f}_m\|^2 \\ \le \; & C_1 \inf_j \left(\frac{1}{n/2} \log \frac{1}{\pi_j} + C_2 E\|\sigma^2 - \hat{\sigma}^2_{j,n/2}\|^2 + \frac{(n/2-1)\, C_2}{n/2} E\|f - \hat{f}_{j,n/2}\|^2 + \frac{C_2}{n/2} E\|f - \hat{f}_{j,n}\|^2 \right). \end{aligned}$$

The bound is slightly better than the earlier one. Thus averagely speaking, $\widetilde{f}_m$, $n/2+1 \le m \le n$, behaves well adaptively. Note that $\widetilde{f}_n$ redefined here when averaged over random permutations gives the estimator $\hat{f}_n$ defined in Section 2. One might expect the above risk bound holds for $\hat{f}_n$ (or $\widetilde{f}_n$) individually since intuitively, $\hat{f}_n$ (or $\widetilde{f}_n$) should behave no worse than $\widetilde{f}_m$ for $n/2+1 \le m \le n-1$ based on more observations. But we do not know whether this is generally true. With either definition of $\widetilde{f}_n$, one can take the simple average of the estimators, i.e., $\frac{1}{n/2} \sum_{m=n/2+1}^{n} \widetilde{f}_m$ and the risk bounds still hold respectively (by virtue of convexity). Since averaging increases the computation, though we do not have an individual risk bound for $\hat{f}_n$, we recommend it to be used in applications. The simulations in Section 4 do show its good performance in several settings.

Based on the above discussion, from Theorem 1, by mixing different candidate procedures, we have a single procedure that shares the advantages of them automatically in terms of the risk.

For the case with multiple candidate error distributions involved in the procedures, let h_f denote the true error distribution and let Δ_f be the sub-collection of the procedures δ_j in Δ for which h_f is being used in the construction of $W_{j,i}$ in (1). We have the following performance bound on ARM for this case.

Theorem 1': Assume that Condition A1 holds and A2 is satisfied for each $h \in \mathcal{H}$. For any given countable collection of estimation procedures $\Delta = \{\delta_j, j \geq 1\}$ and a list of error distributions $\mathcal{H}$, we can construct a single combined estimation procedure δ_* such that

$$R(f; n; \delta_*) \leq C_1 \inf_{j \in \Delta_f} \left(\frac{1}{N_n} \log \frac{1}{\pi_j} + \frac{C_2}{N_n} \sum_{l=n-N+1}^{n} \left(E\|\sigma^2 - \hat{\sigma}^2_{j,l}\|^2 + E\|f - \hat{f}_{j,l}\|^2 \right) \right), \tag{4}$$

where the constant C_1 depends on A and $\overline{\sigma}$, and C_2 depends on $A, \overline{\sigma}/\underline{\sigma}$ and h_f.

It is worth noting that the risk bound involves only those procedures with the right choice of error distribution. Thus at least asymptotically speaking, exploring with a few error distributions will not have much of an adverse effect. Of course, including too many different error distributions will unavoidably reduce the prior weight π_j on the good procedures, which can substantially damage the performance.

6 Risk bounds with homoscedastic errors

Compared to estimating the regression function, there is much less work on estimating the variance function. A recent work in that direction is by Ruppert *et al* (1997), where a local polynomial method is proposed with a theoretical justification. In this section, we present some results on ARM when the variance function is assumed to be a constant. Here a simple estimator of the variance σ^2 is computed for each estimator of the regression function according to its performance in prediction. The data are split into three portions with the first part used for estimating f by each procedure, second portion used for estimating σ^2 and the last part used for measuring the performance of the procedures based on prediction. The algorithm can be very useful when no reliable variance estimator is available for the procedures. We assume that h is known here. The method also applies when there are multiple candidate error distributions as for Theorem 1'.

Three-Stage ARM

Step 0. Randomly permute the order of the observations.

Step 1. Split the data into three parts $Z^{(1)} = (X_i, Y_i)_{i=1}^{n_1}$, $Z^{(2)} = (X_i, Y_i)_{i=n_1+1}^{n_1+n_2}$, and $Z^{(3)} = (X_i, Y_i)_{i=n_1+n_2+1}^{n}$. Let $n_3 = n - n_1 - n_2$.

Step 2. Obtain estimates $\hat{f}_{j,n_1}(x; Z^{(1)})$ of f based on $Z^{(1)}$ for $1 \leq j \leq J$.

Step 3. Estimate the variance σ^2 by

$$\hat{\sigma}^2_j = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} \left(Y_i - \hat{f}_{j,n_1}(X_i) \right)^2.$$

Step 4. For each j, evaluate predictions. For $n_1+n_2+1 \le i \le n$, predict Y_i by $\hat{f}_{j,n_1}(X_i)$. Compute

$$E_j = \frac{\Pi_{i=n_1+n_2+1}^{n} h\left(\frac{(Y_i - \hat{f}_{j,n_1}(X_i))}{\hat{\sigma}_j}\right)}{\hat{\sigma}_j^{n_3}}.$$

Step 5. Compute the weight for procedure δ_j. Let

$$W_j = \frac{E_j}{\sum_{l=1}^{J} E_l}.$$

Step 6. Repeat steps 0-5 $(M-1)$ more times and average the weights over the M random permutations. Let $\hat{W}_j$ denote the weight of procedure δ_j obtained this way. The final estimator is

$$\widetilde{f}_n(x) = \sum_{j=1}^{J} \hat{W}_j \hat{f}_{j,n}(x) \tag{5}$$

We derive below a theoretical property of the three-stage ARM in terms of an average risk as for ARM in the previous section. For each $n_1+n_2+1 \le m \le n-1$, define

$$E_{j,m} = \frac{\Pi_{i=n_1+n_2+1}^{m} h\left(\frac{(Y_i - \hat{f}_{j,n_1}(X_i))}{\hat{\sigma}_j}\right)}{\hat{\sigma}_j^{m-(n_1+n_2)}},$$

$$W_{j,m} = \frac{E_{j,m}}{\sum_{l=1}^{J} E_{l,m}}$$

and

$$\widetilde{f}_m(x) = \sum_{j=1}^{J} W_{j,m} \hat{f}_{j,n_1}(x).$$

We assume σ^2 is upper bounded by a known constant $\overline{\sigma}^2$ here. One can accordingly restrict $\hat{\sigma}_j^2$ in the interval $(0, \overline{\sigma}^2]$.

Theorem 2: Assume the conditions for Theorem 1 hold, and the error distribution h has a finite fourth moment, i.e., $E\varepsilon_i^4 < \infty$. Then the average risk of $\widetilde{f}_m$, $n_1+n_2+1 \le m \le n$, satisfies

$$\frac{1}{n_3}\sum_{m=n_1+n_2+1}^{n} E\|f-\widetilde{f}_m\|^2 \le C_1 \inf_j \left(\frac{1}{n_3}\log\frac{1}{\pi_j} + \frac{C_3}{n_2} + \frac{C_4(n_3-1)}{n_3}E\|f-\hat{f}_{j,n_1}\|^2 + \frac{C_2}{n_3}E\|f-\hat{f}_{j,n_1+n_3}\|^2\right),$$

where the constant C_1 depends on A and $\overline{\sigma}$, C_2 and C_4 depend on A and $\overline{\sigma}/\underline{\sigma}$ and h, and C_3 depends on $E\varepsilon_i^4$ and A.

Remark: As in Section 5, one can take the Cesaro average of the estimators $\widetilde{f}_m$ to get an estimator with individual risk bounded by the same quantity above. But due to much more computation, we recommend the use of $\widetilde{f}_n$ as defined in (5).

The above risk bound quantifies the trade-off among three sources of discrepancies: estimating f using the first n_1 observations, estimating the variance for each procedure based on n_2 observations, and

assessing performance in prediction to assign weights using n_3 observations. A choice with n_1, n_2, and n_3 all of order n (e.g., $n_1 = n_2 = n_3 = n/3$) usually gives the optimal rate of convergence offered by the candidate procedures. Note that except the boundness assumption, no other condition (e.g., continuity) is required on the regression function.

Differently from estimating σ^2 based on the individual models, a common estimator may also be used. For example, for the one-dimensional case, one can use

$$\hat{\sigma}_n^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(Y_{(i+1)} - Y_{(i)} \right)^2,$$

where $Y_{(i)}$ denotes the observed response at the ith smallest X value (see, Rice (1984)). Under a mild smoothness assumption on f, $\hat{\sigma}_n^2$ has squared risk converging at the parametric rate $1/n$, which also does not affect the rate of convergence for estimating the regression function f.

7 Conclusion

In this paper, we proposed an algorithm ARM to combine candidate regression procedures. It is theoretically shown that ARM provides the intended adaptation capability. ARM can be used to combine parametric models and/or nonparametric procedures. It can be very useful when there is difficulty comparing the candidate procedures. Simulation results strongly support the adaptation property in various settings.

8 Proofs of the results

Proof of Theorem 1: The joint density of (X, Y) (with respect to the product measure of P_X and μ) under f and σ^2 is denoted $p_{f,\sigma}(x, y)$, i.e.,

$$p_{f,\sigma}(x, y) = \frac{1}{\sigma(x)} h\left(\frac{y - f(x)}{\sigma(x)} \right).$$

Define

$$q_{n-N+1}(x, y; z^{n-N+1}) = \textstyle\sum_{j \geq 1} \pi_j p_{\hat{f}_{j,n-N+1}, \hat{\sigma}_{j,n-N+1}}(x, y)$$

$$q_{n-N+2}(x, y; z^{n-N+2}) = \frac{\sum_{j \geq 1} \pi_j p_{\hat{f}_{j,n-N+1}, \hat{\sigma}_{j,n-N+1}}(x_{n-N+2}, y_{n-N+2}) p_{\hat{f}_{j,n-N+2}, \hat{\sigma}_{j,n-N+2}}(x, y)}{\sum_{j \geq 1} \pi_j p_{\hat{f}_{j,n-N+1}, \hat{\sigma}_{j,n-N+1}}(x_{n-N+2}, y_{n-N+2})}$$

...

$$q_i(x, y; z^i) = \frac{\sum_{j \geq 1} \pi_j \left(\Pi_{l=n-N+1}^{i-1} p_{\hat{f}_{j,l}, \hat{\sigma}_{j,l}}(x_{l+1}, y_{l+1}) \right) p_{\hat{f}_{j,i}, \hat{\sigma}_{j,i}}(x, y)}{\sum_{j \geq 1} \pi_j \Pi_{l=n-N+1}^{i-1} p_{\hat{f}_{j,l}, \hat{\sigma}_{j,l}}(x_{l+1}, y_{l+1})}$$

...

$$q_n(x, y; z^n) = \frac{\sum_{j \geq 1} \pi_j \left(\Pi_{l=n-N+1}^{n-1} p_{\hat{f}_{j,l}, \hat{\sigma}_{j,l}}(x_{l+1}, y_{l+1}) \right) p_{\hat{f}_{j,n}, \hat{\sigma}_{j,n}}(x, y)}{\sum_{j \geq 1} \pi_j \Pi_{l=n-N+1}^{n-1} p_{\hat{f}_{j,l}, \hat{\sigma}_{j,l}}(x_{l+1}, y_{l+1})}.$$

Since h has mean 0, given x, $q_i(x,y;Z^i)$ has mean $\sum_j W_{j,i}\hat{f}_{j,i}(x)$ in y, where $W_{j,i}$ are defined in (1). Let

$$\hat{g}_n(y|x) = \frac{1}{N}\sum_{i=n-N+1}^{n} q_i(x,y;Z^i).$$

Given x, $\hat{g}_n$ is a convex combination of densities in y of the form $h\left(\frac{y-b}{a}\right)/a$ with random locations and scales depending on the data (but not on knowledge of P_X or f). It can be viewed as an estimator of the conditional density of Y given $X=x$. Note that given x, the mean of $\hat{g}_n(y|x)$ is exactly the estimator $\overline{\hat{f}}_n(x)$ given in (2) in Section 5.

For simplicity in notation, let i_0 denote $n-N+1$ and denote $(x_l,y_l)_{l=i_0+1}^{n+1}$ by $z_{i_0+1}^{n+1}$. Let $\hat{p}_{j,i}(x,y;z^i) = p_{\hat{f}_{j,i},\hat{\sigma}_{j,i}}(x,y)$ for $i\geq 1$ and $j\geq 1$. Let

$$g_j\left(z_{i_0+1}^{n+1}\right) = \Pi_{l=i_0}^{n}\hat{p}_{j,l}(x_{l+1},y_{l+1};z^l).$$

Given z^{i_0} and $x_{i_0+1},...,x_{n+1}$, it is a probability density function in $y_{i_0+1},...,y_{n+1}$. Mixing these densities over different procedures (j), we define

$$g^{(n)}\left(z_{i_0+1}^{n+1}\right) = \sum_{j\geq 1}\pi_j g_j\left(z_{i_0+1}^{n+1}\right).$$

Let $\hat{m}_l(x,y) = q_l(x,y;Z^l)$. Note that $\hat{m}_l$ can be viewed as an estimator of the joint density of (X,Y) with respect to the product measure of $P=P_X$ and μ. The cumulative risk of $\hat{m}_l$ based on Z^l, $i_0\leq l\leq n$ under the Kullback-Leibler (K-L) divergence can be bounded in terms of the individual risks of the original procedures using an idea of Barron (1987) originated in information theory (the idea is also used in Yang and Barron (1999) in deriving minimax rates of convergence for a general function class, and in Yang (2000a) on combining density estimators). Indeed, we have

$$\begin{aligned}
&\textstyle\sum_{l=i_0}^{n} ED(p_{f,\sigma}\,\|\,\hat{m}_l)\\
&= \textstyle\sum_{l=i_0}^{n} E\int p_{f,\sigma}(x,y)\log\frac{p_{f,\sigma}(x,y)}{\hat{m}_l(x,y)}\mu(dy)P(dx)\\
&= \textstyle\sum_{l=i_0}^{n} E\int p_{f,\sigma}(x_{l+1},y_{l+1})\log\frac{p_{f,\sigma}(x_{l+1},y_{l+1})}{\hat{m}_l(x_{l+1},y_{l+1})}\mu(dy_{l+1})P(dx_{l+1})\\
&= E\int \Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})\left(\textstyle\sum_{l=i_0}^{n}\log\frac{p_{f,\sigma}(x_{l+1},y_{l+1})}{q_l(x_{l+1},y_{l+1};Z_1,\ldots,Z_{i_0},z_{i_0+1},\ldots,z_{l+1})}\right)\mu(dy_{i_0+1})...\mu(dy_{n+1})P(dx_{i_0+1})\cdots P(dx_{n+1})\\
&= E\int \Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})\left(\log\frac{\Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})}{g^{(n)}\left(z_{i_0+1}^{n+1}\right)}\right)\mu(dy_{i_0+1})...\mu(dy_{n+1})P(dx_{i_0+1})\cdots P(dx_{n+1}).
\end{aligned}$$

For any f,σ and any $j\geq 1$, since $\log(x)$ is an increasing function, we have

$$\begin{aligned}
&\int \Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})\left(\log\frac{\Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})}{g^{(n)}\left(z_{i_0+1}^{n+1}\right)}\right)\mu(dy_{i_0+1})...\mu(dy_{n+1})P(dx_{i_0+1})\cdots P(dx_{n+1})\\
\leq\ &\int \Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})\left(\log\frac{\Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})}{\pi_j g_j\left(z_{i_0+1}^{n+1}\right)}\right)\mu(dy_{i_0+1})...\mu(dy_{n+1})P(dx_{i_0+1})\cdots P(dx_{n+1})\\
=\ &\log\frac{1}{\pi_j}+\int \Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})\log\frac{\Pi_{l=i_0}^{n}p_{f,\sigma}(x_{l+1},y_{l+1})}{g_j(z_{i_0+1}^{n+1})}\mu(dy_{i_0+1})...\mu(dy_{n+1})P(dx_{i_0+1})\cdots P(dx_{n+1}).
\end{aligned}$$

The last term above can be bounded in terms of risks of the estimators produced by strategy δ_j. Indeed, as earlier but going backwards, we have

$$E \int \Pi_{l=i_0}^{n} p_{f,\sigma}(x_{l+1}, y_{l+1}) \log \frac{\Pi_{l=i_0}^{n} p_{f,\sigma}(x_{l+1}, y_{l+1})}{g_j\left(z_{i_0+1}^{n+1}\right)} \mu(dy_{i_0+1}) \ldots \mu(dy_{n+1}) P(dx_{i_0+1}) \cdots P(dx_{n+1})$$

$$= \sum_{l=i_0}^{n} ED(p_{f,\sigma} \parallel \hat{p}_{j,l}).$$

Now

$$\begin{aligned} D(p_{f,\sigma} \quad \parallel \quad \hat{p}_{j,l}) &= \int \left(\int \frac{1}{\sigma(x)} h\left(\frac{y - f(x)}{\sigma(x)}\right) \log \frac{\frac{1}{\sigma(x)} h\left(\frac{y-f(x)}{\sigma(x)}\right)}{\frac{1}{\hat{\sigma}_{j,l}(x)} h\left(\frac{y-\hat{f}_{j,l}(x)}{\hat{\sigma}_{j,l}(x)}\right)} \mu(dy) \right) P(dx) \\ &= \int \left(\int h(y) \log \frac{h(y)}{\frac{\sigma(x)}{\hat{\sigma}_{j,l}(x)} h\left(\frac{\sigma(x)}{\hat{\sigma}_{j,l}(x)} y + \frac{f(x) - \hat{f}_{j,l}(x)}{\hat{\sigma}_{j,l}(x)}\right)} \mu(dy) \right) P(dx) \\ &\leq B \left(\int \left(\frac{\hat{\sigma}_{j,l}(x)}{\sigma(x)} - 1 \right)^2 P(dx) + \int \left(\frac{f(x) - \hat{f}_{j,l}(x)}{\sigma(x)} \right)^2 P(dx) \right) \\ &\leq \frac{B}{\underline{\sigma}^2} \left(\int \left(\sigma(x) - \hat{\sigma}_{j,l}(x)\right)^2 P(dx) + \int \left(f(x) - \hat{f}_{j,l}(x) \right)^2 P(dx) \right), \end{aligned}$$

where the second equality follows from a simple linear transformation in integration, the first inequality follows from the assumption A2 on h with B depending on h, A, and $\overline{\sigma}/\underline{\sigma}$, and the second inequality follows from our assumption A1 on the variance functions. Thus we have shown

$$\sum_{l=i_0}^{n} ED(p_{f,\sigma} \parallel \hat{m}_l) \leq \log \frac{1}{\pi_j} + \frac{B}{\underline{\sigma}^2} \sum_{l=i_0}^{n} \left(E \| \sigma^2 - \hat{\sigma}_{j,l}^2 \|^2 + E\left(\| f - \hat{f}_{j,l} \|^2 \right) \right). \tag{6}$$

For $\hat{g}_n(y|x) = (1/N) \sum_{l=i_0}^{n} \hat{m}_l(x, y)$, by convexity of the K-L divergence in its second argument, we have

$$ED(p_{f,\sigma} \parallel \hat{g}_n) \leq \frac{1}{N} \sum_{l=i_0}^{n} ED(p_{f,\sigma} \parallel \hat{m}_l).$$

Since the above inequality (6) holds for all j, minimizing over j, we have

$$ED(p_{f,\sigma} \parallel \hat{g}_n) \leq \inf_j \left\{ \frac{1}{N} \log \frac{1}{\pi_j} + \frac{B}{\underline{\sigma}^2 N} \sum_{l=i_0}^{n} E \parallel \sigma^2 - \hat{\sigma}_{j,l}^2 \parallel^2 + \frac{B}{\underline{\sigma}^2 N} \sum_{l=i_0}^{n} E \parallel f - \hat{f}_{j,l} \parallel^2 \right\}.$$

Let $d_H^2(p, q) = \int \left(\sqrt{p} - \sqrt{q}\right)^2 d\nu$ denote the squared Hellinger distance between the densities f and g with respect to a measure ν. Since the squared Hellinger distance is upper bounded by the K-L divergence, the above risk bound implies

$$Ed_H^2(p_{f,\sigma}, \hat{g}_n) \leq \inf_j \left\{ \frac{1}{N} \log \frac{1}{\pi_j} + \frac{B}{\underline{\sigma}^2 N} \sum_{l=i_0}^{n} E \parallel \sigma^2 - \hat{\sigma}_{j,l}^2 \parallel^2 + \frac{B}{\underline{\sigma}^2 N} \sum_{l=i_0}^{n} E \parallel f - \hat{f}_{j,l} \parallel^2 \right\}.$$

Now for each x, $\hat{g}_n$ naturally gives an estimator of $f(x)$ by taking the mean value of $\hat{g}_n$ with respect to y at the given x. For this estimator, we have

$$\begin{aligned}
& \left(\int y p_{f(x),\sigma(x)}(x,y)\mu(dy) - \int y\hat{g}_n(y|x)\mu(dy) \right)^2 \\
= & \left(\int y \left(p_{f(x),\sigma(x)}(x,y) - \hat{g}_n(y|x) \right) \mu(dy) \right)^2 \\
= & \left(\int y \left(\sqrt{p_{f(x),\sigma(x)}(x,y)} + \sqrt{\hat{g}_n(y|x)} \right) \left(\sqrt{p_{f(x),\sigma(x)}(x,y)} - \sqrt{\hat{g}_n(y|x)} \right) \mu(dy) \right)^2 \\
\leq & \int y^2 \left(\sqrt{p_{f(x),\sigma(x)}(x,y)} + \sqrt{\hat{g}_n(y|x)} \right)^2 \mu(dy) \int \left(\sqrt{p_{f(x),\sigma(x)}(x,y)} - \sqrt{\hat{g}_n(y|x)} \right)^2 \mu(dy) \\
\leq & 2 \left(\int y^2 p_{f(x),\sigma(x)}(x,y)dy + \int y^2 \hat{g}_n(y|x)\mu(dy) \right) \int \left(\sqrt{p_{f(x),\sigma(x)}(x,y)} - \sqrt{\hat{g}_n(y|x)} \right)^2 \mu(dy) \\
= & 2 \left(f^2(x) + \sigma^2(x) + \int y^2 \hat{g}_n(y|x)\mu(dy) \right) d_H^2 \left(p_{f(x),\sigma(x)}(x,\cdot), \hat{g}_n(\cdot \mid x) \right),
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality, and $d_H^2\left(p_{f(x),\sigma(x)}(x,\cdot), \hat{g}_n(\cdot \mid x)\right)$ in the last equality is the Hellinger distance between $p_{f(x),\sigma(x)}(x,\cdot)$ and $\hat{g}_n(\cdot \mid x)$ in terms of y given x. As mentioned earlier, given x, $\hat{g}_n(y|x)$ is a convex combination of the densities of the location-scale family $h\left((y-b)/a\right)/a$ in h with means $\hat{f}_{j,l}(x)$, $j \geq 1$ and $i_0 \leq l \leq n$, and standard deviations $\hat{\sigma}_{j,l}(x)$. Under Assumption A1, the regression estimators are bounded between $-A$ and A and the variance estimators are bounded above by $\overline{\sigma}^2$. Thus $\int y^2 \hat{g}_n(y|x)\mu(dy) \leq A^2 + \overline{\sigma}^2$. It follows that

$$\int \left(f(x) - \int y\hat{g}_n(y|x)\mu(dy) \right)^2 P(dx) \leq 4(A^2 + \overline{\sigma}^2) \int d_H^2 \left(p_{f(x),\sigma(x)}(x), \hat{g}_n(\cdot \mid x) \right) P(dx).$$

Together with $\int y\hat{g}_n(y|x)\mu(dy) = \overline{\hat{f}}_n(x)$, we have

$$\begin{aligned}
& E \int (f(x) - \hat{f}_n(x))^2 P(dx) \\
\leq & 4(A^2 + \overline{\sigma}^2) E d_H^2(p_{f,\sigma}, \hat{g}_n) \\
\leq & 4(A^2 + \overline{\sigma}^2) \inf_j \left\{ \frac{1}{N} \log \frac{1}{\pi_j} + \frac{B}{N\underline{\sigma}^2} \sum_{l=i_0}^{n} E \parallel \sigma^2 - \hat{\sigma}_{j,l}^2 \parallel^2 + \frac{B}{N\underline{\sigma}^2} \sum_{l=i_0}^{n} E \parallel f - \hat{f}_{j,l} \parallel^2 \right\}.
\end{aligned}$$

The second conclusion of Theorem 1 on the average risk follows similarly by working with lower bounding $D(p_{f,\sigma} \parallel \hat{m}_l)$, $i_0 \leq l \leq n$ instead of $ED(p_{f,\sigma} \parallel \hat{g}_n)$. This completes the proof of Theorem 1.

Proof of Theorem 2: Similarly to the derivation for Corollary 1, Theorem 1 implies

$$\frac{1}{n_3} \sum_{m=n_1+n_2+1}^{n} E \parallel f - \widetilde{f}_m \parallel \leq C_1 \inf_j \left(\frac{1}{n_3} \log \frac{1}{\pi_j} + C_2 E(\sigma^2 - \hat{\sigma}_j^2)^2 + \frac{C_2\,(n_3-1)}{n_3} E\|f - \hat{f}_{j,n_1}\|^2 + \frac{C_2}{n_3} E\|f - \hat{f}_{j,n_1+n_3}\|^2 \right).$$

It remains to bound the risks of the variance estimators. By definition of $\hat{\sigma}_j^2$, we have

$$\hat{\sigma}_j^2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} (\varepsilon_i + f(X_i) - \hat{f}_{j,n_1}(X_i))^2$$

$$= \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\varepsilon_i^2 + \frac{2}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\varepsilon_i(f(X_i)-\hat{f}_{j,n_1}(X_i)) + \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(f(X_i)-\hat{f}_{j,i}(X_i))^2.$$

Expanding squares and observing that most cross-product terms disappear after taking expectation due to independence, we have

$$\begin{aligned}
& E(\sigma^2-\hat{\sigma}_j^2)^2 \\
= \;& E\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(\varepsilon_i^2-\sigma^2)\right)^2 + E\left(\frac{2}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\varepsilon_i(f(X_i)-\hat{f}_{j,n_1}(X_i))\right)^2 \\
& + E\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(f(X_i)-\hat{f}_{j,n_1}(X_i))^2\right)^2 + 2E\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}(\varepsilon_i^2-\sigma^2)\right)\left(\frac{2}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\varepsilon_i(f(X_i)-\hat{f}_{j,n_1}(X_i))\right) \\
= \;& \frac{Var(\varepsilon_i^2)}{n_2} + \sum_{i=n_1+1}^{n_1+n_2}\frac{4}{n_2^2}E\left(\varepsilon_i(f(X_i)-\hat{f}_{j,n_1}(X_i))\right)^2 + \left(E(f(X_{n_1+1})-\hat{f}_{j,n_1}(X_{n_1+1}))^2\right)^2 \\
& + \frac{1}{n_2}Var\left((f(X_{n_1+1})-\hat{f}_{j,n_1}(X_{n_1+1}))^2\right) + \frac{4}{n_2}\left(E\varepsilon_i^3\right)E\left(f(X_{n_1+1})-\hat{f}_{j,n_1}(X_{n_1+1})\right) \\
= \;& \frac{Var(\varepsilon_i^2)}{n_2} + \frac{4\sigma^2}{n_2}E\|f-\hat{f}_{j,n_1}\|^2 + \left(E\|f-\hat{f}_{j,n_1}\|^2\right)^2 + \frac{1}{n_2}Var\left((f(X_{n_1+1})-\hat{f}_{j,n_1}(X_{n_1+1}))^2\right) \\
& + \frac{4}{n_2}\left(E\varepsilon_i^3\right)E\left(f(X_{n_1+1})-\hat{f}_{j,n_1}(X_{n_1+1})\right) \\
\leq \;& \frac{Var(\varepsilon_i^2)}{n_2} + \frac{4\sigma^2}{n_2}E\|f-\hat{f}_{j,n_1}\|^2 + 4A^2E\|f-\hat{f}_{j,n_1}\|^2 + \frac{4A^2}{n_2}E\|f-\hat{f}_{j,n_1}\|^2 + \frac{4\left(E\varepsilon_i^4\right)^{3/4}}{n_2}\sqrt{E\|f-\hat{f}_{j,n_1}\|^2}.
\end{aligned}$$

For the last inequality, we used the boundness assumption on the regression function and the estimators. Note that when σ^2 is upper bounded by $\overline{\sigma}^2$, restricting $\hat{\sigma}_j^2$ in the interval $(0,\overline{\sigma}^2]$ certainly does not increase the risk $E(\sigma^2-\hat{\sigma}_j^2)^2$. The conclusion of the theorem then follows from above together with the boundness assumption on f. This completes the proof of Theorem 2.

9 Acknowledgments

The author is grateful to Hyung-Woo Kim for conducting most of the simulations in this work. He also thanks a referee, an associate editor and the editor for their very helpful comments.

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? *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., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields,* **113** 301-413.

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

[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 advice? *Journal of the ACM*, **44**, 427-485.

[11] Cesa-Bianchi, N. and Lugosi, G. (1999) On prediction of individual sequences. Accepted by *Ann. Statistics.*

[12] Clemen, R.T. (1989) Combining forecasts: a review and annotated bibliography. *Intl. J. Forecast.*, **5**, 559-583.

[13] Cleveland, W.S. (1993) *Visualizing Data*. Hobart Press, Summit, NJ.

[14] Foster, D.P. and Vohra, R.V. (1993) A randomization rule for selecting forecasts. *Operations Research*, **41**, 704-709.

[15] Hart, J.D. (1997) *Nonparametric Smoothing and Lack-of-fit Tests*, Springer-Verlag.

[16] Haussler, D., Kivinen, J. and Warmuth, M.K. (1998) Sequential prediction of individual sequences under general loss functions. *IEEE Trans. on Information Theory*, **44**, 1906-1925.

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

[18] Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Comp.*, **3**, 79-87.

[19] Jiang, W. and Tanner, M.A. (1999) Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Ann. Statistics,* **27**, 987-1011.

[20] Jiang, W. and Tanner, M.A. (2000) On the asymptotic normality of hierarchical mixtures-of experts for generalized linear models. *IEEE Trans. on Information Theory*, **46**, 1005-1013.

[21] Jordan, M.I. and Jacobs, R.A. (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Comp.*, **6**, 181-214.

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

[23] Kivinen, J. and Warmuth, M.K. (1999) Averaging expert predictions. *Eurocolt 99.*

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

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

[26] Loader, C.R. (1999) Bandwidth selection: Classical or plug-in? *Ann. Statist.*, **27**, 415-438.

[27] Lugosi, G. and Nobel, A. (1999) Adaptive model selection using empirical complexities. Accepted by *Ann. Statistics.*

[28] Merhav, N. and Feder, M. (1998) Universal prediction. *IEEE Trans. on Information Theory*, **44**, 2124-2147.

[29] Olkin, I. and Spiegelman, C.H. (1987) A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.*, **82**, 858-865.

[30] Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

[31] Ruppert, D., Wand, M.P., Holst, U., and Hössjer, O. (1997) Local polynomial variance-function estimation. *J. Amer. Statist. Assoc.*, **39**, 262-273.

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

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

[34] Stone, M. (1974) Cross-validatory Choice and Assessment of Statistical Predictions (with Discussion). *J. Roy. Statist. Soc., Ser.* B, **36**, 111-147.

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

[36] Vovk, V.G. (1998) A game of prediction with expert advice. *Journal of Computer and System Sciences*, **56** 153-173.

[37] Wolpert, D. (1992) Stacked generalization. *Neural Networks*, **5**, 241-259.

[38] Yang, Y. (1999a) Model selection for nonparametric regression. *Statistica Sinica*, **9**, 475-499.

[39] Yang, Y. (1999b) Regression with multiple candidate models: selecting or mixing? Technical Report No. 8, Department of Statistics, Iowa State University.

[40] Yang, Y. (2000a) Mixing strategies for density estimation. *Ann. Statististics*, **28**, 75-87.

[41] Yang, Y. (2000b) Combining Different Procedures for Adaptive Regression. *Journal of Multivariate Analysis*, **74**, 135-161.

[42] Yang, Y. (2000c) Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica*, **10** (to appear).

[43] Yang, Y. and Barron, A.R. (1998) An asymptotic property of model selection criteria. *IEEE Trans. on Information Theory*, **44**, 95-116.

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