Model Combining in Factorial Data Analysis

Lihua Chen Yuhong Yang

312 Snedecor Hall

Department of Statistics and Statistical Laboratory

Iowa State University

Ames, IA 50011-1210

Email: lchen@iastate.edu yyang@iastate.edu

October, 2002

Abstract

ANOVA is a statistical tool commonly used to study factor effects. Usually a model is selected through hypothesis testing or according to some model selection criteria. Both approaches have difficulties in finding the best model. When the goal is to estimate the cell means rather than to select a model, we proposed an algorithm, ARM, to convexly combine the candidate models. Simulation and data examples demonstrate the advantage of combining over selection when there is high instability in model selection. A theoretical risk bound on the combined estimator is also obtained.

Keywords: ANOVA models, Model selection instability, Model combining

1 Introduction

Analysis of variance (ANOVA) is a commonly used statistical technique for studying the relationship between a response variable and one or more explanatory factors. Specifically, the ANOVA technique is used to get information about the main effects and joint effects of the factors, to test various contrasts and to identify the "best" combination of factors.

Traditionally, analysis of variance has focused on the identification of significant factors. This reflects the design of experiment perspective, which aims to find out which factors significantly affect the response variable. Usually, when one compares a few models, one uses the ANOVA tables to eliminate the insignificant effects and factors. Under the normality assumption on the errors, elegant theoretical results on the decomposition of sum of squares provide powerful tools for comparing two nested models in the traditional hypothesis testing framework.

In applications of ANOVA, however, more often than not one can not confidently narrow the list of potential candidate models to a few models. In such a case, if testing is sequentially done, as it is often done in applications, little can be said on the overall performance of the whole procedure. In addition, another drawback associated with the hypothesis testing approach to comparing models is the arbitrary choice of the test size. It is unclear how such a choice is related to the performance of the selected model for the purpose of parameter estimation and prediction. Also there is a gray area where one may not feel comfortable in drawing definite conclusions.

The approach of using a model selection criterion is less subjective and perhaps better guided for overall performance. Various model selection criteria can be used for this purpose such as AIC (Akaike (1973)), cross validation (Allen (1974), Stone (1974)) and BIC (Schwarz (1978)). A problem common to data-driven methods of model selection is the potential for large instability in searching for the best model. Here "instability" is interpreted as the uncertainty in identifying the best model (in terms of a statistical risk of interest). Often a small change or perturbation of the data results in the selection of a quite different model. As a consequence, the prediction or estimation based on the selected model may have high variability.

Many efforts have been directed toward reducing the instability associated with model selection. Breiman (1996b) proposed *bagging* which involves generating multiple bootstrap versions of an estimator and then averaging them into a stabilized estimator. Empirical evidence showed an advantage for bagging in terms of estimation accuracy. Another approach to reducing 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)), though attention has been mainly given to regression. Buckland, Burnham and Augustin (1997) proposed a plausible model weighting method based on the values of a model selection criterion (e.g., AIC).

It is well known that one disadvantage of combining is the difficulty in interpreting the estimator. However, in the ANOVA setting, since the cell means can be uniquely decomposed into main effects and interaction effects under a certain restriction on the parameters, factor effects can be analyzed effectively

through a good estimation of cell means by combining. Of course estimation of cell means is of interest in itself in many applications. For this reason, estimation of cell means in ANOVA models through combining becomes our main interest in this paper. The contributions of this work are:

1. We propose instability measures for model selection. Whenever model selection is involved, such measures are valuable for providing a reasonable sense of how trustworthy inference based on the final selected model is. We advocate the report of such measures in statistical applications in which model selection is involved. Our simulation and data examples show that often the objective of finding the correct or the best model is unrealistic.

2. When there is significant model selection instability, we propose the method ARM for combining (mixing) candidate models. A theoretical result is derived for the combining method.

3. We compare the performance of a hypothesis testing approach, model selection methods, and ARM in simulations and data examples.

ARM was proposed in Yang (2001a) for regression with random covariates. The methodology is modified in this work to deal with the ANOVA setting.

The paper is organized as follows. In Section 2, we set up the problem of interest. Section 3 proposes several approaches to measuring instability associated with model selection. In Section 4, we present the ARM algorithms. Intensive simulation results are given in Section 5. Concluding remarks are in Section 6. A risk bound to theoretically characterize its performance and the proofs of the theoretical results on ARM are given in an appendix.

2 Problem Setup

Suppose there are m ($m \geq 2$) factors with levels $I_1, \cdots, I_m$ ($I_1, \cdots, I_m \geq 2$) respectively. Consider a balanced factorial design with J replicates. Let $Y_{i_1 \cdots i_m, j} = \mu_{i_1 \cdots i_m} + \epsilon_{i_1 \cdots i_m, j}$, where $Y_{i_1 \cdots i_m, j}$ is the jth observation in cell $i_1 \cdots i_m$, $\mu_{i_1 \cdots i_m}$ is the mean response at that cell and $\epsilon_{i_1 \cdots\ i_m, j}$ are independent Gaussian errors with mean 0 and unknown variance σ^2 ($\sigma^2 > 0$). ANOVA concerns how the cell means $\mu_{i_1 \cdots i_m}$ depend on the factors and also the estimation of main factor and interaction effects.

To estimate the cell means vector $\mu = \{\mu_{i_1 \cdots i_m}\}$, K plausible models are considered:

$$Y = \mu^{(k)} + \varepsilon,$$

where $Y = \{y_{i_1 \cdots i_m, j}\}$ is the data vector, $\epsilon = \{\epsilon_{i_1 \cdots i_m, j}\}$ is the error vector and for each $k \in \{1, ..., K\}$, and $\mu^{(k)}$ is a family of mean functions. For example, $k = 1$ may be the independence model that

includes only the main effects and $k = 2$ may be the model including all the main effects and all the 2 way interactions.

In this paper, the comparison of estimators will be based on the average mean square error. Let $\hat{\mu}$ be an estimator of μ based on the data, the risk is

$$R(\mu, \hat{\mu}) = \frac{1}{N} \sum_{i_1=1}^{I_1} \cdots \sum_{i_m=1}^{I_m} E(\mu_{i_1 \cdots i_m} - \widehat{\mu}_{i_1 \cdots i_m})^2 = \frac{1}{N} E(\| \mu - \hat{\mu} \|^2)$$

where N is the total number of cells in the model and the expectation is taken with respect to the randomness of the errors under the true model. In this paper, under the Gaussian errors, we will use the least square estimators (which is also MLE) to estimate the cell means.

3 Instability in Model Selection

As mentioned in the introduction, uncertainty in model selection in general has been well recognized. Model averaging techniques have been proposed as an alternative. However, there is no clear understanding of when model averaging methods outperform model selection in the estimation of cell means of factorial data.

Evidence from other contexts indicates that model selection may be more appropriate under some circumstances than others. We would expect our proposed method to perform better in cases where model selection is less appropriate. A formal measure that could quantify the appropriateness of model selection given a set of data could serve as a guide to understanding the properties of combining and selection in this investigation, and as a potential guide in applications to help decide whether to choose selecting or combining.

We propose such a measure based on a criterion of internal consistency. When a model selection technique initially chooses one model but chooses a different model when conditions are changed slightly, we say the model selection technique displays instability. We consider three ways in which conditions can be changed slightly. The data could be perturbed, as in measurement error, the data could be reduced, as in moving from a larger to a smaller experiment, or data could be redrawn from the same stable process as in tests repeated over time. We call the three instability measures corresponding to these three forms of slight change *perturbation instability,sequential instability*, and *parametric bootstrap instability* respectively.

The three model selection methods considered here are AIC, BIC, and methods based on hypothesis tests. The first two methods can be applied to our data sets directly, but we have to make a choice to

address the third. Model selection based on hypothesis testing is approached in a variety of ways and often there are several ways to implement each approach, each potentially leading to different outcomes.

In a balanced design, F tests can be performed to study the factor and interaction effects. The most commonly used test size is .05, which is the default value in many statistical packages. The F test plays an important role in model comparison and selection. In the simplest case, it can be used to compare two nested models. One common way of applying F tests to study factor effects starts with the full model and obtains the ANOVA table from the full model. Then all the terms that are not significant at 0.05 level are dropped. The remaining terms constitute the selected model. We will call this approach the ANOVA method. Because it is common, well understood, and gives unequivocal results suitable to computational evaluation, we will use it as the representative of hypothesis test based model selection.

3.1 Some data sets

Six data sets will be used to demonstrate the proposed instability measures. We briefly describe the data sets below.

Data set 1 (Neter, et al. (1996, p. 942)):

A 2^3 experiment. Each treatment has three replicates.

Data set 2 (Vardeman S (2001, p. 191))

A 2^3 experiment. Each treatment has three replicates.

Data set 3 (Montgomery D (1996, p.341)):

A 2^3 experiment. Each treatment has two replicates.

Data set 4 (Garcia-Diaz A and Phillips D.T (1995, p.218)):

A 2^3 experiment. Each treatment has two replicates.

Data set 5 (Montgomery D (1996, p.345)):

A 2^4 experiment. Each treatment has two replicates.

Data set 6 (McLean R.A and Anderson V.L (1984, p.7)):

A 2^4 experiment. Each treatment has two replicates.

3.2 Parametric Bootstrap Instability

The idea of parametric bootstrap can be naturally used to measure model selection instability. Consider a model selection method. The selected model is used to get the estimated cell means $\widehat{\mu}_{i_1 \cdots i_m}$ and the estimate of the error variance $\widehat{\sigma}^2$. Then in each cell, J observations are generated from $N(\hat{\mu}_{i_1 \cdots i_m}, \hat{\sigma}^2)$

and the selection method is applied to the new data. The procedure is repeated a large number of times (say 100) and the relative frequency with which it chooses a different model is recorded. If the frequency is high, the selection process is unstable. The results of this measure for the data sets are summarized in table 1.

	ANOVA	AIC	BIC
Data set 1	0.36	0.34	0.36
Data set 2	0.14	0.20	0.06
Data set 3	0.22	0.32	0.33
Data set 4	0.41	0.54	0.52
Data set 5	0.58	0.54	0.59
Data set 6	0.75	0.66	0.52

Table 1: Parametric Bootstrap Instability of the Data Sets

3.3 Sequential Instability

Sequential instability examines the consistency of selection at different data sizes. We expect that removing a small proportion of the data shouldn't make much difference if a procedure is stable. In the balanced design, we randomly remove 1 observation from each cell and apply the same model selection procedure to the remaining data. The relative frequency with which a different model is chosen in 100 replications is recorded.

We apply this approach only to the data sets with at least 3 replicates. For the cases with only 2 replicates, removing one observation per cell implies reduction of the samples size by one half, which may have quite different statistical behavior. The results are summarized in table 2.

	ANOVA	AIC	BIC
Data set 1	0.33	0.45	0.44
Data set 2	0	0.14	0.07
Data set 3	0.21	0.35	0.42

Table 2: Sequential Instability of the Data Sets

Regarding the validity of the sequential instability, naturally, one may wonder whether high frequencies of choosing a different model are due to model selection instability or due to the reduction of sample size. Does the best model change when we change the data size by 1/3? A simulation to address this concern is given below.

Let's consider the model $y = a+b+c+ab$, with parameters taking values $a_1 = 0.75, b_1 = -0.50, c_1 =$

0.25, and $ab_{11} = 0.125$ with $\sigma^2 = 1$. Three replicates in each cell are generated from this model. Fit the data on all candidate models and calculate the mean square error based on the difference between estimated and true cell means. Randomly delete 1 data point in each cell and calculate the mean square error on the remaining data. The model with the minimum average mean square error after 100 repetitions of this procedure is identified as the best model in terms of the statistical risk.

In both the full and the reduced data sets, the model $y = a + b + c$ was identified as the best model. Apply AIC and BIC to the data and we found when the data size changed, AIC and BIC chose different models 53 and 45 times respectively. In both data sizes, they did not choose the best model in more than 90 times out of 100 times.

This showed that while the data size changed, the model with the smallest risk remained the same, but model selection methods displayed much instability.

3.4 Perturbation Instability

The perturbation approach to measuring instability involves perturbing each data point by a small amount and reselecting to see if the model selected changes. For each data point y, a perturbed data point is generated from $N(y, \tau\hat{\sigma}^2)$, where τ is the scaler factor and $\hat{\sigma}^2$ is the estimated variance from the model selected based on the original data. The model selection method is repeated on the perturbed data. The procedure is repeated 100 times and we record the relative frequency with which a different model is chosen.

	τ=0.2	τ=0.4	τ=0.6	τ=0.8	τ=1
ANOVA	0	0.08	0.34	0.48	0.58
AIC	0	0.13	0.35	0.37	0.63
BIC	0	0.15	0.33	0.49	0.69

Table 3: Perturbation Instability of Data Set 1

	τ=0.2	τ=0.4	τ=0.6	τ=0.8	τ=1
ANOVA	0	0	0	0	0.03
AIC	0	0	0.01	0.02	0.03
BIC	0	0.01	0.01	0.01	0.01

Table 4: Perturbation Instability of Data Set 2

	τ=0.2	τ=0.4	τ=0.6	τ=0.8	τ=1
ANOVA	0	0	0	0.12	0.22
AIC	0	0.15	0.25	0.27	0.44
BIC	0	0.10	0.36	0.42	0.67

Table 5: Perturbation Instability of Data Set 3

	τ=0.2	τ=0.4	τ=0.6	τ=0.8	τ=1
ANOVA	0.39	0.51	0.55	0.73	0.77
AIC	0.22	0.49	0.56	0.62	0.79
BIC	0.23	0.51	0.57	0.64	0.84

Table 6: Perturbation Instability of Data Set 4

3.5 Analysis of Results

Data set 2 had the smallest instability by all three measures for all three model selection techniques. Data set 3 was next, though perturbation instability was higher at $\tau = 0.4$ for AIC and was more unstable at $\tau = 0.6$ for BIC.

Parametric bootstrap instability ranked the remaining data sets, from lowest instability, as 4, 5, and 6 in all methods except for BIC, where the order of 5 and 6 was reversed. Thus the over all ranking is 2,3,1,4,5,6, with the last three methods ranked by only two measures. This order is roughly consistent with the idea that more complex models should be harder to estimate - the three factor models were more stable than the four factor models, and that more data should make estimation easier - the two replication cases showed more instability. The anomaly in this ranking is data set 1, a three factor data set with three replications that proved less stable than than one with only two replications. This could be accounted for by a low noise level in data set three or a high one in data set one.

The instability of ANOVA in data set 4 was lower than that of the other two methods, though not enough to reverse ranking. The instability of BIC for data set 5 was about the same as that of the other two methods. But in data set 6, BIC had much lower instability than the other methods, which in turn were far from each other in instability. Data set 6 thus appeared to be a data set of highest instability and a data set in which there was the most variance among the three methods.

In perturbation instability, data set 4 was the most consistently ranked at all levels of τ and was the most stable of the last three. ANOVA ranked the three data sets most consistently, while BIC was most inconsistent at different levels of τ. BIC had most inconsistent in its ranking for the most unstable data set. BIC again had about the same amount of instability for data set 4, but the instability it had at different levels of τ were very wide spread for data set 6. The rankings largely agreed at each level of τ. While reported as a table of values at present, this measure will be refined as single slope rather than points at various τ levels. The reversals may be artifacts of the small number of replications presently

	τ=0.2	τ=0.4	τ=0.6	τ=0.8	τ=1
ANOVA	31	51	56	77	87
AIC	32	58	77	83	87
BIC	0	25	47	73	76

Table 7: Perturbation Instability of Data Set 5

	τ=0.2	τ=0.4	τ=0.6	τ=0.8	τ=1
ANOVA	0.37	0.53	0.70	0.89	0.96
AIC	0.54	0.76	0.72	0.77	0.88
BIC	0	0.12	0.45	0.71	0.86

Table 8: Perturbation Instability of Data Set 6

used in calculating the measure. This too will be refined for this as well as the other measures making them all more precise. The characterization is none the less fairly clear and consistent. The measures of instability, while broadly consistent on ranking of the data sets, are not identical. They vary over model selection methods and they vary in magnitude across data sets. Thus they appear to be informative about the different methods on the same data set and about the character of different data sets.

The combining method to be described next will be applied to the preceding data sets and analyzed with respect to the results just summarized. The results suggest it the combining method should perform better on data set 6 than on data set 2 and that data sets 3 and 4 should be in between.

4 An algorithm for model combining

The ARM algorithm proposed by Yang uses a portion of the data to fit each candidate model and the other portion of the data to evaluate the performance of each candidate model. The candidate models are weighted according to their performance in the evaluation stage and combined to give the ARM estimator. In this paper, we adapt the ARM algorithm to the special case of a balanced factorial design. The details of the algorithm are as follows:

Algorithm

- *Step 1.* Randomly permute the order of the J observations within each cell.
- *Step 2.* Split the data into 2 parts. In each cell, the first part has J_1 observations, the second part has J_2 observations. The data in each cell are split in the same proportion to maintain the balanced design. Note $J = J_1 + J_2$. The first part of data contains $n_1 = J_1 N$ observations and is denoted by $Z^{(1)}$, the second part contains $n_2 = J_2 N$ observations and is denoted by $Z^{(2)}$.

- *Step 3.* For each candidate model $k = 1, 2...K$, obtain $\hat{\mu}^{(k)} = \hat{\mu}^{(k)}_{n_1}$ by least square method based on $Z^{(1)}$. Obtain the estimate of the variance $\hat{\sigma}_k = \hat{\sigma}^2_{k,n_1}$ from the same set of data.

- *Step 4.* Assess the performance of the models using $Z^{(2)}$, the remaining part of the data $Z^{(2)}$, according to the overall measure of discrepancy $D_k = \sum_{i_1=1}^{I_1} \cdots \sum_{i_m=1}^{I_m} \sum_{j=J_1+1}^{J} (Y_{i_1 \cdots i_m, j} - \hat{\mu}_{i_1 \cdots i_m})^2$.

- *Step 5.* Assign each model k the weight

$$W_k = \frac{\left(\hat{\sigma}^2_k\right)^{-n_2/2} \exp\left(-\hat{\sigma}_k^{-2} D_k/2\right)}{\sum_{l=1}^{L} \left(\hat{\sigma}^2_l\right)^{-n_2/2} \exp\left(-\hat{\sigma}_l^{-2} D_l/2\right)}.$$

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

- *Step 6.* Repeat steps 1-5 $M - 1$ times and average the weights over the M random permutations. Let $\hat{W}_k$ be the weight of the k th model obtained this way. Compute the convex combination of estimators produced by the models:

$$\tilde{\mu}_{i_1 \cdots i_m} = \sum_{k=1}^{K} \hat{W}_k \hat{\mu}^{(k)}_{i_1 \cdots i_m}$$

Remarks:

1. If we put the uniform prior on the models and take the estimates of μ and σ based on the first part of the data as the true values of the models, then W_k may be interpreted as the posterior probability of model k after observing the second part of the data. Our motivation and justification, however, is not necessarily Bayesian. Note that ARM is not a formal Bayes procedure. In particular, no averaging over parameters is performed.

2. Note that $\widehat{\mu}_n$ depends on all the estimators from the candidate models. When the uncertainty of finding the best model is small, the lack of interpretability is a serious drawback of a model combining method. When the uncertainty is large, however, since the selected model is not trustworthy, insisting on interpretability is not appropriate.

3. Note for the estimation of σ^2, if we take model dependent variance estimation method using $\hat{\sigma}^2_k = SSE_k/(n - p_k)$, where SSE_k is the sum of squared residuals, n is the sample size and p_k is the number of parameters in the model k. We will encounter difficulty in estimating σ^2 for the full model if there is only one observation in the first part of the data. In this case, we can borrow the variance estimation from the other models. One reasonable approach is to borrow the variance

estimate from the next largest model. This is the approach we adopted in the simulations. Another approach is to estimate the variance using the full data. Also the variance can be estimated by pooled sample variances across all the cells.

5 Empirical Studies

In this section, we will compare the performance of ARM with that of some model selection methods under a certain assessment criterion. Comparisons are made with both simulations and real data examples. With simulated data, the assessment criterion is the risk discussed is Section 2. With the real data sets, the assessment criterion is the square prediction error. The model selection methods considered here include AIC, BIC, ANOVA, and CV. The CV we consider chooses a model in the following way: Randomly spare one data point in each cell as test data. Denote the test data by $(y_{i_1\cdots i_m})$. For each model, get the estimate of the cell means $\hat{\mu}_{i_1\cdots i_m}$ based on the remaining data. Calculate the square prediction error by $\sum_{i_1=1}^{I_1}\cdots\sum_{i_m=1}^{I_m}(y_{i_1\cdots i_m}-\hat{\mu}_{i_1\cdots i_m})^2$. We repeat the procedure 50 times to average out the splitting effect. The model with the minimum average square prediction error is selected.

The simulations start with the selection of a true model. The true cell means $\mu_{i_1\cdots i_m}$ are calculated according to the model. In each cell J observations are generated from $N(\mu_{i_1\cdots i_m},\sigma^2)$. We generate a large number of (say L) data sets from the same true model and use the average of the square errors from these L replications as a Monte Carlo approximation of risk (average mean square error) for AIC, BIC, ANOVA, CV and ARM respectively. With each replication, the data are permuted M times to average out instability in splitting which occurs in both CV selection and model combining.

For selection and combining, the candidates are the 5 possible models in 2 factor design, the candidates ar e the 19 possible models in the 3 factor design, and the candidates are the 167 possible models in 4 factor case (we include the null model as a candidate). Note that as usual, only hierarchical models are considered as candidate models.

We consider several settings. First, some fixed 2, 3 and 4 factor models and some randomly generated models are analyzed. Results on some data examples follow the simulations. In all these settings, each factor has two levels.

5.1 Fixed Models

In each case, 100 data sets are generated from one specified true model and the average mean square error are based on these 100 replications. With each replication, the data are permuted 50 times to average out splitting variability. The simulation results are in table 9-13. The number in the parenthesis

is the standard error of the average square error.

Case 1 Two factors: Data are generated from the model $y = a + b + ab$, with $a_1 = 0.75, b_1 = -0.68, ab_{11} = 0.29$. Each cell has 2 replicates.

	AIC	BIC	CV	ARM
$\sigma^2 = 0.5$	0.304	0.309	0.333	0.302
	(0.022)	(0.022)	(0.022)	(0.016)
$\sigma^2 = 1$	0.604	0.621	0.585	0.394
	(0.035)	(0.035)	(0.033)	(0.023)
$\sigma^2 = 1.5$	0.923	0.929	0.891	0.605
	(0.049)	(0.049)	(0.052)	(0.041)

Table 9: Comparing Model Selection to Combining with 2 Factor Models

Since we had only 5 available candidates models, we expected the model selection procedures didn't have much difficulty in identifying the best model. But ARM still exhibited advantages when we increased the noise level to high values.

Case 2 Three factors: Data are generated from $y = a + b + c + ab$ with $a_1 = 0.75, b_1 = 0.68, c_1 = 0.29, ab_{11} = 0.12$. Each cell has 3 replicates. We can see as the noise level increased, the advantage of

	AIC	BIC	CV	ARM
$\sigma^2 = 0.5$	0.134	0.140	0.136	0.113
	(0.008)	(0.008)	(0.007)	(0.006)
$\sigma^2 = 1$	0.274	0.316	0.290	0.209
	(0.017)	(0.022)	(0.019)	(0.013)
$\sigma^2 = 1.5$	0.440	0.516	0.449	0.302
	(0.030)	(0.032)	(0.035)	(0.019)

Table 10: Comparing Model Selection to Combining with 3 Factor Models

combining also increased. At the three noise levels, ARM achieved a reduction in risk over the best one of AIC, BIC and CV by 13.3%, 20.9% and 26.2% respectively.

case 3 Three factors: In this case, we keep the same model as in case 2 and the same parameter values except change $ab_{11} = 0.32$. Each cell has three replicates. We increase the magnitude of the interaction term so it not vague any more. When the model does not contain obviously weak interaction terms, ARM still has a big advantage as in the previous case. That may be due to the fact that there still exist relatively weak terms in the model. The reduction in risk by ARM is 13.2%, 16.8% and 26.2% respectively.

Case 4 Four factors: Data are generated from the model $y = a + b + c + d + bc + cd$, with $\mu_0 = 0, a_1 = 0.75, b_1 = -0.46, c_1 = -0.25, d_1 = 0.29, bc_{11} = 0.12, cd_{11} = -0.30$. Each cell has 2 replicates.

In this case, we included a weak $b * c$ interaction term. ARM showed substantial advantage at all the noise levels. Again, the advantage increased when the noise level increased from 0.5 to 1.5. ARM

	AIC	BIC	CV	ARM
$\sigma^2 = 0.5$	0.171	0.186	0.167	0.145
	(0.010)	(0.009)	(0.009)	(0.007)
$\sigma^2 = 1$	0.327	0.369	0.343	0.272
	(0.017)	(0.023)	(0.019)	(0.015)
$\sigma^2 = 1.5$	0.454	0.532	0.470	0.335
	(0.028)	(0.031)	(0.028)	(0.019)

Table 11: Comparing Model Selection to Combining with 3 Factor Models

	AIC	BIC	CV	ARM
$\sigma^2 = 0.5$	0.166	0.200	0.178	0.130
	(0.008)	(0.011)	(0.009)	(0.005)
$\sigma^2 = 1$	0.390	0.419	0.399	0.240
	(0.018)	(0.016)	(0.016)	(0.010)
$\sigma^2 = 1.5$	0.529	0.529	0.528	0.322
	(0.024)	(0.024)	(0.025)	(0.016)

Table 12: Comparing Model Selection to Combining with 4 Factor Models

achieved reduction in risk by 21.7%, 38.4 and 39.0% respectively.

Case 5 Four factors: Data are generated from the model $y = a + b + c + d$, with the parameters taking the same values as in case 4. Each cell has 2 replicates. It might be interesting to examine the

	AIC	BIC	CV	ARM
$\sigma^2 = 0.5$	0.134	0.149	0.140	0.094
	(0.008)	(0.008)	(0.008)	(0.005)
$\sigma^2 = 1$	0.325	0.330	0.337	0.202
	(0.019)	(0.016)	(0.018)	(0.009)
$\sigma^2 = 1.5$	0.399	0.444	0.418	0.245
	(0.023)	(0.023)	(0.023)	(0.013)

Table 13: Comparing Model Selection to Combining with 4 Factor Models

frequency with which AIC, BIC chose the true model. The following table shows the frequency with which the true model was chosen in 100 replications. Note as noise level increased, the frequency with which AIC and BIC chose the true model decreased greatly. That may help explain why they performed much worse than ARM in these scenarios.

The true model contained no obviously weak terms, but ARM still had an advantage even at low noise level. The reduction in risk by ARM was 29.9%, 37.8% and 38.6% respectively. It seems when the true model involves more factors and terms, ARM has potentially bigger advantages (Compare to the 2 factor case). That makes intuitive sense as a more complicated model is more difficult to identify.

	AIC	BIC
$\sigma^2 = 0.5$	18	23
$\sigma^2 = 1$	7	6
$\sigma^2 = 1.5$	4	2

Table 14: Frequency in Selecting True Model with AIC, BIC

5.2 Random Models

In order to show that the above results hold in more general cases, we consider random models in this section. A random model is generated in the following way: The list of all possible models for that number of factors is partitioned into groups that have the same number of terms. Call this the model size. A model size is then selected at random. One model from that size class is then selected at random. Parameters for the main effects and the intercept are generated from uniform (-1, 1). The parameters for the interaction terms are generated from uniform (-1, 1) or uniform (-0.3, 0). A noise level of $\sigma = 1$ is used to generated the data.

Case 1 Three factors: All the parameters are assigned values from uniform (-1, 1). 50 models are generated by the above mentioned procedure. For each model, $K = 20$ replications are made with $M = 50$ permutations in each replication to smooth splitting variability in each replication. The box plot in **Figure 1** is based on the fifty average square errors from the fifty models.

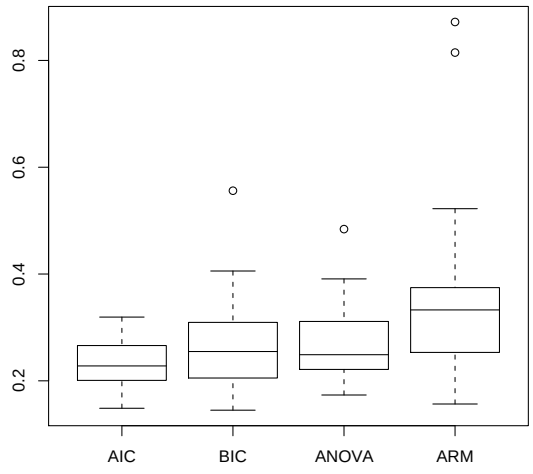

Figure 1. Comparing Model Selection to Combing with Random Simulation

Case 2 Three factors: The only difference between case 2 and case 1 is that the parameters of the interactions are generated from uniform (-0.3, 0) and therefore are weaker. The results are shown in **Figure 2**.

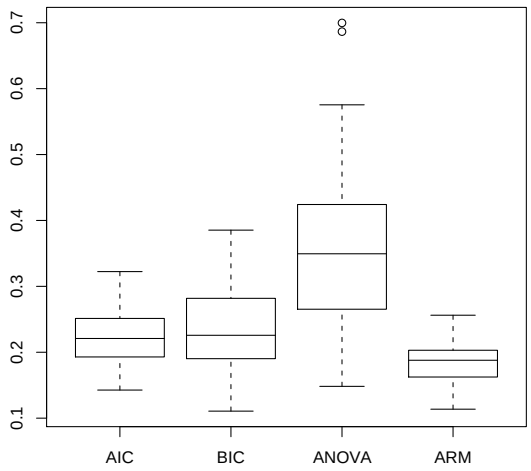

Figure 2. Comparing Model Selection to Combing with Random Simulation

Case 3 Four factors: The parameters of the main terms are generated from uniform (-1, 1). The parameters of the interactions are generated from uniform (-0.3, 0). The box plot in **Figure 3** is based on the simulation results from 50 models.

The simulation results from random models reaffirmed our earlier findings: As the true model got more complicated (i.e.,include more terms), the advantage of ARM increased. When the true model contained weak interaction terms (with parameters values generated from uniform (-1, 1)) which hamper the ability of model selection to identify the true model, the gain from ARM also increased substantially.

5.3 Data Examples

In this section, we apply ARM to the data sets that were used in section three on instability.

In data sets 1 through 3, which have three replicates in each cell, we randomly spare one data point in each cell as test data and calculate the square prediction error for each method. We repeat the procedure 100 times to average out splitting variability. The average square prediction error based on the 100 replications for each method is in table 15. The number in the parenthesis is the standard error

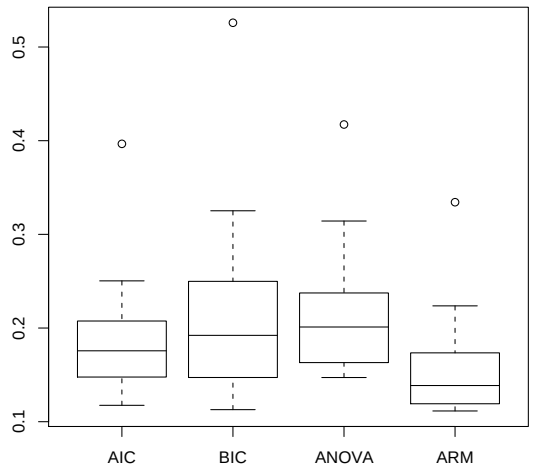

Figure 3. Comparing Model Selection to Combining with Random Simulation

of average square prediction error.

In data sets 4 through 6, each treatment has two observations. We spare one data point from $N/2$ randomly selected cells as test data. As the remaining observations constitute an unbalanced design, we do not consider the ANOVA method here. The average square prediction errors based on 100 replications are in table 16.

	ANOVA	AIC	BIC	ARM
Data set 1	15.761	15.761	15.761	13.464
	(0.472)	(0.472)	(0.472)	(0.579)
Data set 2	1.731	2.384	2.384	2.216
	(0.053)	(0.127)	(0.127)	(0.095)
Data set 3	59.125	54.573	56.974	53.164
	(1.58)	(1.48)	(1.44)	(1.37)

Table 15: Comparing Model Selection to Combining with Data 1 through 3

For the first four data sets, the performance of ARM increases as the instability of the data set increases. After the most stable data set by all measures, data set 2, ARM starts to perform better than the model selection methods. Its advantage in risk increases from 2.6% to 15% and 18.8%. However, the advantage stops increasing for the four factor models holding steady at 12.5% and 12.3%. While the precise relationship between the instability measures and ARM performance is still to be worked out, the evidence here is consistent with an instability threshold past which model combining becomes advantageous.

	AIC	BIC	ARM
data set 4	40.027	41.537	32.501
	(1.261)	(1.259)	(1.313)
data set 5	0.040	0.042	0.035
	(0.003)	(0.004)	(0.002)
data set 6	15.166	15.457	13.295
	(0.527)	(0.461)	(0.401)

Table 16: Comparing Model Selection to Combining with Data 4 through 6

6 Concluding Remarks

Estimation of cell means is of great importance in experimental designs. Estimation through model selection often leads to unsatisfactory results due to instability associated with model selection. In this paper, we presented several approaches to measuring the instability associated with model selection. Due to the drawbacks of model selection, we proposed the use of a combining method, ARM, to convexly combine the candidate models to try to reduce variability and improve estimation accuracy. Simulation results and data examples demonstrated the advantage of combining over model selection in this setting.

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 very small, splitting the data may cause problems in estimation (e.g., having more parameters than the number of observations).

Based on the studies we have done so far, we found:

- Model selection usually has large instability in searching for the best model in ANOVA modeling involving three or more factors. The instability tends to increase when more factors are present.
- When a model selection criterion has no difficulty in identifying the best model, model selection usually outperforms ARM.
- ARM has a substantial advantage over model selection when there is high uncertainty in model selection.
- In three and four factor cases, it is computationally feasible to combine all possible models and simply combining all models produce quite satisfactory results. For applications, one can use a model selection method and/or graph inspection to screen out models that are obviously inappropriate to reduce the list of models to be combined.

7 Theory and proof

7.1 A risk bound for ARM

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

$$\sup_{k\geq 1} \| \widehat{\mu}^{(k)} - \mu \|_\infty \leq \sqrt{\tau}\sigma$$

Condition 2: The variance estimators $\widehat{\sigma}_k^2$ are not too far away from the true value: there exist constants $0 < \xi_1 \leq 1 \leq \xi_2 < \infty$ such that

$$\xi_1 \leq \frac{\widehat{\sigma}_k^2}{\sigma^2} \leq \xi_2$$

with probability one for all $k \geq 1$. The above conditions are satisfied if the estimation function and the error variance are upper and lower bounded by known constants and the estimators are accordingly restricted to that range. 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. First let us simplify the notation. Split the data into two parts with J_1 and J_2 observations in each cell. Hence the second part of the data contains $n_2 = J_2 N$ observations in total. Stack these n_2 observations into one vector $Y = (y_1, y_2 \cdots y_{n_2})$. The data are stacked in the following order: 1) The observations in the same cell are stacked together. 2) For the cell order, we let the last factor change fastest, and let the first factor change slowest. Denote the mean of the cell where y_i belongs to by μ_i. For $i = n_1 + 1$, let $W_{k,i} = 1/K$ and for $n_1 + 1 \leq i \leq n$, let

$$W_{k,i} = \frac{(\hat{\sigma}_k)^{-(i-n_1-1)} \exp\left(-\frac{1}{2\hat{\sigma}_k^2}\sum_{l=n_1+1}^{i-1}(Y_l - \widehat{\mu}_l)^2\right)}{\sum_{k=1}^{K}(\hat{\sigma}_k)^{-(i-n_1-1)} \exp\left(-\frac{1}{2\hat{\sigma}_k^2}\sum_{l=n_1+1}^{i-1}(Y_l - \widehat{\mu}_l)^2\right)}.$$

Then define

$$\widetilde{W}_k = \frac{1}{n_2}\sum_{i=n_1+1}^{n} W_{k,i}.$$

and let

$$\widetilde{\mu}_n = \sum_{k=1}^{K} \widetilde{W}_k \widehat{\mu}_{n_1}^{(k)} \tag{1}$$

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

$$E \parallel \widetilde{\mu}_n - \mu \parallel^2 \leq (1 + \xi_2 + 9\tau/2) \inf_{j \geq 1} \left(\frac{4\sigma^2 \log K}{n} + \frac{1}{\xi_1} E \parallel \widehat{\mu}_{n_1}^{(k)} - \mu \parallel^2 + C(\xi_1, \xi_2) E(\widehat{\sigma}_{k,n_1}^2 - \sigma^2)^2 \right).$$

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

Remarks:

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. Note as far as we know, there is no non-asymptotic risk bound on estimators based on model selection.

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 $\widehat{\mu}_{n_1}^{(k)}$ 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 eventually 2.

Note that the estimator $\widetilde{\mu}_n$ in Theorem 1 as defined by (1) is not exactly the same as $\widetilde{\mu}_n$ 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.

7.2 Proof of theorem 1:

Proof: Let $n_1 = J_1 N$ and $n_2 = J_2 N$ be the sizes of the estimation and evaluation portions of the data. Let $\widehat{\mu}^{(k)}$ denote $\widehat{\mu}_{n_1}^{(k)}$ and $\widehat{\sigma}_k^2$ denote $\widehat{\sigma}_{k,n_1}^2$ for $k \geq 1$. For simplicity in notation, in this proof, we drop the bold face format for a vector. Let

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

and

$$\begin{aligned} q^{n_2} &= \frac{1}{K} \sum_{k=1}^{K} \Pi_{i=n_1+1}^{n} \frac{1}{\sqrt{2\pi\widehat{\sigma}_k^2}} \exp\left(-\frac{1}{2\widehat{\sigma}_k^2} (y_i - \widehat{\mu}_i^{(k)})^2 \right) \\ &= \frac{1}{K} \sum_{k=1}^{K} \frac{1}{(2\pi\widehat{\sigma}_k^2)^{n_2/2}} \exp\left(-\frac{1}{2} \sum_{i=n_1+1}^{n} \frac{(y_i - \widehat{\mu}_i^{(k)})^2}{\widehat{\sigma}_k^2} \right). \end{aligned}$$

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

$$\begin{aligned}\log\left(p^{n_2}/q^{n_2}\right) &\leq \log\left(\frac{\left(2\pi\sigma^2\right)^{-n_2/2}\exp\left(-\frac{1}{2}\sum_{i=n_1+1}^{n}\frac{(y_i-\mu_i)^2}{\sigma^2}\right)}{\frac{1}{K}\left(2\pi\widehat{\sigma}^2_{k^*}\right)^{-n_2/2}\exp\left(-\frac{1}{2}\sum_{i=n_1+1}^{n}\frac{(y_i-\widehat{\mu}_i^{(k)})^2}{\widehat{\sigma}^2_{k^*}}\right)}\right) \\ &= \log K + \frac{1}{2}\sum_{i=n_1+1}^{n}\left(\log\frac{\widehat{\sigma}^2_{k^*}}{\sigma^2}+\frac{(y_i-\widehat{\mu}_i^{(k)})^2}{\widehat{\sigma}^2_{k^*}}-\frac{(y_i-\mu_i)^2}{\sigma^2}\right). \qquad (2)\end{aligned}$$

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

$$E_{n_1}\left(\log\frac{\widehat{\sigma}^2_{k^*}}{\sigma^2}+\frac{(y_{i_1\ldots i_m,j}-\widehat{\mu}^{(k)}_{i_1\ldots i_m})^2}{\widehat{\sigma}^2_{k^*}}-\frac{(y_{i_1\ldots i_m,j}-\mu_{i_1\ldots i_m})^2}{\sigma^2}\right)=\frac{\parallel\widehat{\mu}^{(k)}-\mu\parallel^2}{\widehat{\sigma}^2_{k^*}}+\frac{\sigma^2}{\widehat{\sigma}^2_{k^*}}-1-\log\frac{\sigma^2}{\widehat{\sigma}^2_{k^*}}. \quad (3)$$

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

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sqrt{2\pi\widehat{\sigma}^2_k}}\exp\left(-\frac{1}{2\widehat{\sigma}^2_k}(y_{n_1+1}-\widehat{\mu}^{(k)}_{n_1+1})^2\right) \\ &\times\frac{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{(\sqrt{2\pi\widehat{\sigma}^2_k})^2}\exp\left(-\frac{1}{2\widehat{\sigma}^2_k}[(y_{n_1+1}-\widehat{\mu}^{(k)}_{n_1+1})^2+(y_{n_1+2}-\widehat{\mu}^{(k)}_{n_1+2})^2]\right)}{\sum_{k=1}^{K}\frac{1}{\sqrt{2\pi\widehat{\sigma}^2_k}}\exp\left(-\frac{1}{2\widehat{\sigma}^2_k}(y_{n_1+1}-\widehat{\mu}^{(k)}_{n_1+2})^2\right)} \\ \times\quad&\cdots\cdots\times\frac{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{(\sqrt{2\pi\widehat{\sigma}^2_k})^{n_2}}\exp\left(-\sum_{i=n_1+1}^{n}\frac{1}{2\widehat{\sigma}^2_k}(y_i-\widehat{\mu}^{(k)}_i)^2\right)}{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{(\sqrt{2\pi\widehat{\sigma}^2_k})^{n_2-1}}\exp\left(-\sum_{i=n_1+1}^{n-1}\frac{1}{2\widehat{\sigma}^2_k}[(y_i-\widehat{\mu}^{(k)}_i)^2]\right)}\end{aligned}$$

Let $p_i=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)$ and
$g_i=\sum_{k=1}^{K}W_{k,i}\frac{1}{\sqrt{2\pi\widehat{\sigma}^2_k}}\exp\left(-\frac{(y_i-\widehat{\mu}_i^{(k)})^2}{2\widehat{\sigma}^2_k}\right)$ for $n_1+1\leq i\leq n$. It follows by the definition of $W_{k,i}$ that $\log\left(p^{n_2}/q^{n_2}\right)=\sum_{i=n_1+1}^{n}\log\left(\frac{pi}{gi}\right)$. Together with (2) and (3), under the assumptions on the data, we have

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

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

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

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 2001). 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 $0 < \sigma_p^2 < \ nfty$ and $0 < \sigma_g^2 < \infty$ respectively. Then

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

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

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

Together with (4), we have

$$\sum_{i=n_1+1}^{n} E\left(\frac{\left(\widehat{s}_i - \mu_i\right)^2}{\sigma^2\left(2(1+\xi_2) + 9\tau\right)}\right) \leq \log K + \frac{n_2}{2} E\left(\frac{\parallel \widehat{\mu}^{(k)} - \mu \parallel^2}{\widehat{\sigma}_{k^*}^2} + \frac{\sigma^2}{\widehat{\sigma}_{k^*}^2} - 1 - \log\frac{\sigma^2}{\widehat{\sigma}_{k^*}^2}\right).$$

By convexity, we have

$$E\left(\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n} \widehat{s}_i\right) - \mu_i\right)^2 \leq \frac{1}{n_2}\sum_{i=n_1+1}^{n} E\left(\widehat{s}_i - \mu_i\right)^2.$$

Note that $\frac{1}{n_2}\sum_{i=n_1+1}^{n} \widehat{s}_i = \widetilde{\mu}_n$. Thus,

$$E \parallel \widetilde{\mu}_n - \mu \parallel^2 \leq \sigma^2\left(2(1+\xi_2) + 9\tau\right)\left(\frac{\log K}{n_2} + \frac{1}{2}E\left(\frac{\parallel \widehat{\mu}^{(k)} - \mu \parallel^2}{\widehat{\sigma}_{k^*}^2} + \frac{\sigma^2}{\widehat{\sigma}_{k^*}^2} - 1 - \log\frac{\sigma^2}{\widehat{\sigma}_{k^*}^2}\right)\right).$$

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 $c_{x_0} = \frac{x_0 - 1 - \log x_0}{(x_0-1)^2}$. Together with the fact that the above inequality holds for every k^*, under Condition 2, it follows

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

where $C(\xi_1, \xi_2) = \frac{1/\xi_2 - 1 + \log \xi_2}{\xi_1^2(1/\xi_2 - 1)^2}$. The conclusion then follows. This completes the proof of Theorem 1.

Acknowledgments This research was supported by the United States National Science Foundation CAREER Grant DMS0094323.

References

Allen, D.M. (1974), The relationship between variable selection and Data Augmentation and a Method for Prediction. *Technometrics*,**16**,125-127.

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle. *Proc. 2nd Int. Symp. Info. Theory*, 267-281

Breiman, L. (1996a) Stacked regression. *Machine learning*,**24**,49-64.

Breiman, L.(1996b) Bagging predictors.*Machine 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.

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

Garcia-Diaz A., and Phillips D.T. (1995) Principles of experimental design and analysis

Mclean R.A., and Anderson V.L. (1984) Applied factorial and fractional designs

Montgomery D.C. (1997) Design and analysis of experiments

Neter J., Kutner M.H., Nachtsheim C.J., and Wasserman W. (1996) Applied linear statistical models

Schwartz, G.(1978) Estimating the Dimension of a Model. *Ann. Statistics*,**6**, 461-464 Stone, M.(1974) Cross-Validation choice and Assessment of Statistical Prediction. *J. Amer. Statist. Assoc.*,**Ser B, 36**, 111-147.

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

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

Yang,Y.(2002) Regression with multiple candidate models:selecting or combining? *Statistica Sinica*

Yang,Y., and Zou,H. (2002) Combining times series models for forecasting *International Journal of Forecasting*