

Model Combining in Factorial Data Analysis

Lihua Chen*

Department of Mathematics, The University of Toledo

Panayotis Giannakouros

Department of Economics, University of Missouri–Kansas City

Yuhong Yang

School of Statistics, University of Minnesota

Abstract

We study the properties of a model combining method, ARM (Adaptive Regression by Mixing), in the ANOVA framework. We propose model instability measures as a guide to the appropriateness of model combining in applications. We further systematically investigate the relationship between ARM performance and the underlying model structure. We propose an approach to evaluating the importance of factors based on the combined estimates. A theoretical risk bound on the combined estimator is also obtained.

Key words: model combining, model selection instability, ANOVA, adaptive regression by mixing

1 Introduction

A now well known problem common to model selection is the potential for large instability in searching for the best model. By instability we mean the uncertainty in identifying the best model, which in this paper we will take to be best in terms of a statistical risk of interest. Often a small change or

* Corresponding author.

Email addresses: lihua.chen@utoledo.edu (Lihua Chen), poti@potis.org (Panayotis Giannakouros), yyang@stat.umn.edu (Yuhong Yang).

perturbation of the data results in the selection of a quite different model (Breiman, 1996). Inferences based on the selected model are then not reliable.

Model averaging (or combining) is a natural approach to accounting for model uncertainty. In this paper, we investigate a model averaging technique, ARM, proposed by Yang (2001). Compared to the Bayesian model averaging approach which weights the models by their posterior probability (see, e.g. Hoeting et al. (1999) for a review of the general methodology and computational issues), Yang's method does not require the specification of a prior distribution and allows the derivation of theoretical risk bounds for the estimators.

When the selected model is not trustworthy, the estimates from combining are more reliable. However, model averaging is not automatically better than model selection. Model combining is more complicated and computationally expensive, and may even entail sacrificing accuracy. Combined estimators are also difficult to interpret.

The previous work on model averaging has usually been limited to demonstrating the advantage of combining through a few examples without exploring when model combining is more appropriate. We attempt to address this issue in the present work. To that end, we propose instability measures as a guide to help us decide between combining and selection.

To gain more insight into the properties of combining and selection, we explore ARM in the specific context of the ANOVA framework. This is a context with a number of interesting challenges in which model combining does not appear to have been extensively explored. ARM has not been applied to this context. The closest application of model combining we have found is a Bayesian method for generalized linear models developed by Raftery (1996).

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. Model selection is particularly useful as it gives the factor effects directly. Therefore a way to avoid unnecessary model combining is desirable in real applications.

We examine the properties of ARM in the ANOVA setting primarily through an investigation of risk for estimated cell means. We explore the relationship between several proposed instability measures and ARM performance. We also systematically explore the relationship between ARM performance and underlying model structure. Furthermore, given the most common purpose of factorial data analysis, we propose a method for evaluating factors based on the estimated cell means through combining when model selection is not appropriate.

The paper is organized as follows: In section 2, we set up the problem of in-

terest. In section 3, we investigate several approaches to measuring instability associated with model selection. In section 4, we present the ARM algorithm for factorial data. We investigate how to apply ARM and the properties of combining and selection through some data examples and simulations in section 5. Concluding remarks are in section 6. A theoretical risk bound for the ARM estimator and the proof are given in an appendix.

2 Problem Setup

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

To estimate the cell mean vector $\boldsymbol{\mu} = \{\mu_{i_1 \dots i_\Phi}\}$, K plausible models are considered:

$$Y_{i_1 \dots i_\Phi, j} = \mu_{i_1 \dots i_\Phi}^{(k)} + \epsilon_{i_1 \dots i_\Phi, j},$$

where for each $k \in \{1, \dots, K\}$, $\boldsymbol{\mu}^{(k)} = \{\mu_{i_1 \dots i_\Phi}^{(k)}\}$ is a family of mean vectors. 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 two way interactions.

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

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \frac{1}{N} \sum_{i_1=1}^{I_1} \dots \sum_{i_\Phi=1}^{I_\Phi} E(\mu_{i_1 \dots i_\Phi} - \hat{\mu}_{i_1 \dots i_\Phi})^2$$

where $N = \prod_{j=1}^{\Phi} I_j$ is the total number of cells and the expectation is taken with respect to the randomness of the errors under the true model. In this paper, under Gaussian errors, for a given model we will use the least squares estimators (which are also MLE) to estimate the cell means. Throughout this paper, we will impose Σ restrictions on the parameters of the main effects and interactions (i.e., $\sum a_i = 0, \sum b_i = 0$ and so on).

3 Instability in Model Selection

In this part, we study some measures that help us understand when combining outperforms model selection.

Evidence from other contexts indicates that model selection may be more appropriate under some circumstances than others. We would expect model combining to perform better in cases where model selection is less appropriate. A 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, and as a potential guide in applications to help decide whether to choose selecting or combining.

We propose using measures based on criteria of internal consistency. When a model selection technique initially chooses one model but chooses a different model when conditions are changed slightly, we say that 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 data generating 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 (Akaike, 1973), BIC (Schwarz, 1978), and a method based on hypothesis testing. In the ANOVA framework, AIC and BIC can be expressed as

$$\begin{aligned} \text{AIC} &= n \log(RSS/n) + 2p + \text{constant}, \\ \text{BIC} &= n \log(RSS/n) + \log(n)p + \text{constant}, \end{aligned}$$

where RSS is the sum of squared residuals, n is the sample size and p is the number of parameters in the model. The size of p depends on the number of main factor and interactions effects. For example, including a factor with level I_j adds $I_j - 1$ parameters to the model, including an interaction term between two factors with levels I_j and I_k respectively adds $(I_j - 1)(I_k - 1)$ parameters and so on.

The first two methods can be applied to data sets directly, but we have to make a choice to address the third. We will adopt one common approach of studying factor effects: start with the full model and obtain the ANOVA table. 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. Even though diagnosis and iterations are often used in real application, for suitability of automatic 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. The data sets were selected by looking through textbooks and on-line data repositories. We searched until we had an adequate number of data sets that met the criteria of balanced design, three or four factors, two factor levels per factor, and at least two replicates. We briefly describe the data sets below.

Data set 1 (Neter et al., 1996, p. 942): A 2^3 experiment with three replicates.

Data set 2 (Vardeman and Jobe, 2001, p. 191): A 2^3 experiment with three replicates.

Data set 3 (Montgomery, 1997, p. 341): A 2^3 experiment with three replicates.

Data set 4 (Garcia-Diaz and Phillips, 1995, p. 218): A 2^3 experiment with two replicates.

Data set 5 (Montgomery, 1997, p. 345): A 2^4 experiment with two replicates.

Data set 6 (McClean and Anderson, 1984, p. 7): A 2^4 experiment with two replicates.

3.2 *Parametric Bootstrap Instability*

The idea of bootstrapping can be naturally used to measure model selection instability (e.g. Diaconis and Efron, 1983). We here focus on parametric bootstrap. Consider a model selection method. The selected model is used to get the estimated cell means $\hat{\mu}_{i_1 \dots i_\Phi}$ and the estimate of the error variance $\hat{\sigma}^2$. Then in each cell, J observations are generated from $N(\hat{\mu}_{i_1 \dots i_\Phi}, \hat{\sigma}^2)$ and the selection method is applied to the new data. The procedure is repeated a large number of times (say 1000) and the relative frequency with which it chooses a model different from the original selected model is recorded. If the frequency is high, we cannot be confident about the selected model. The results are summarized in Table 1. For data set 6, about 70% of the time, AIC and ANOVA would choose a different model. For such a case, clearly the selected model cannot be trusted as the true or the best model.

	Parametric Bootstrap			Sequential			Perturbation		
	ANOVA	AIC	BIC	ANOVA	AIC	BIC	ANOVA	AIC	BIC
Data set 1	0.387	0.372	0.366	0.344	0.482	0.482	0.331	0.408	0.316
Data set 2	0.165	0.201	0.113	0.000	0.167	0.105	0.002	0.005	0.002
Data set 3	0.192	0.289	0.271	0.180	0.333	0.487	0.041	0.380	0.238
Data set 4	0.446	0.586	0.584	-	-	-	1.342	1.128	1.211
Data set 5	0.580	0.678	0.520	-	-	-	0.717	1.311	1.716
Data set 6	0.716	0.689	0.455	-	-	-	1.538	1.653	0.461

Table 1

Measures of Parametric Bootstrap Instability, Sequential Instability and Perturbation Instability of the 6 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 should not make much difference if a procedure is stable. In the balanced design, we randomly remove 1 observation from each cell and reselect using the remaining data. The relative frequency with which a different model is chosen in 1000 replications is recorded.

We apply this approach only to the data sets with at least three replicates. For the cases with only two 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 1.

3.4 Perturbation Instability

The technique of perturbation was used by Breiman (1996) to demonstrate instability and to obtain a stabilized estimator. Our usage of perturbation is for measuring instability. Here the perturbation approach involves perturbing each data point by a small amount and re-selecting 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 scalar 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 1000 times and the relative frequency with which a different model is chosen is recorded. We let τ vary from 0.1 to 0.5 in increment of 0.1. Intuitively, the relative frequency increases as τ gets larger. We regress the relative frequency versus τ through the origin to get the slope. A high slope implies the selection method

is unstable. Table 1 records the slopes of each method for the data sets.

3.5 Analysis of Results

The ranking of data sets in terms of the instability measures varied quite a bit. Most of the data sets could, however, be roughly characterized as stable, unstable, or intermediate. All the three model selection approaches displayed smallest instability on data set 2 by all three measures. Data set 3 was next. Data sets 4, 5 and 6 which were two replicate data sets were on the whole less stable than the first three data sets, which were all three replicate data sets. We expect the instability to depend on sample size, number of factors, noise level as well as the true coefficients.

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 diversity of the data sets in terms of the instability measures suggests that they should illustrate the performance of the combining method under a wide range of conditions. The results suggest that the combining method should perform better on data sets 1, 4, 5 and 6 than on data sets 2 and 3.

4 Combining Factorial Models

Yang (2001) proposed a method named ARM (adaptation regression by mixing) for combining models in the context of regression with random design. It uses a portion of the data to fit each candidate model and the other portion of the data to evaluate the performance of each model. The models are weighted according to their performance in the evaluation stage and combined to give the ARM estimator. The present ANOVA setting is oriented to substantially different problems from the regression setting in Yang. Taking into account the features of the ANOVA setting different from those of the random design regression setting, we propose the following method to combine ANOVA models.

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 $n_1 = J_1N$ observations and is denoted by $Z^{(1)}$, the second part contains $n_2 = J_2N$ observations and is denoted by

$Z^{(2)}$.

- *Step 3.* For each candidate model $k = 1, 2, \dots, K$, obtain $\hat{\boldsymbol{\mu}}^{(k)} = \hat{\boldsymbol{\mu}}_{n_1}^{(k)}$ by least squares method based on $Z^{(1)}$. Obtain the estimate of the variance $\hat{\sigma}_k^2 = \hat{\sigma}_{k, n_1}^2$ from the same part of the data.
- *Step 4.* Assess the performance of the models using $Z^{(2)}$, the remaining part of the data, according to the overall measure of discrepancy $D_k = \sum_{i_1=1}^{I_1} \cdots \sum_{i_\Phi=1}^{I_\Phi} \sum_{j=J_1+1}^J (Y_{i_1 \dots i_\Phi, j} - \hat{\mu}_{i_1 \dots i_\Phi}^{(k)})^2$.
- *Step 5.* Assign each model k the weight

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

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

- *Step 6.* Repeat steps 1-5 additional $M - 1$ times. Let \hat{W}_k be the averaged weight of the k th model over these M permutations. Compute the convex combination of estimators by:

$$\tilde{\boldsymbol{\mu}} = \sum_{k=1}^K \hat{W}_k \hat{\boldsymbol{\mu}}^{(k)}.$$

Remarks:

- (1) Note that $\tilde{\boldsymbol{\mu}}$ depends on the estimators from all the candidate models.
- (2) For the estimation of σ^2 , one may choose a model dependent variance estimation method using $\hat{\sigma}_k^2 = RSS_k / (n - p_k)$, where RSS_k is the sum of squared residuals, n is the sample size and p_k is the number of parameters in 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. We did not find any substantial difference among these approaches in our empirical investigations.

5 Empirical Studies

In this section, we will compare the performance of ARM with that of some model selection methods in real and simulated data sets. With simulated data, the assessment criterion is the risk discussed in section 2. In the real data sets, the assessment criterion is the squared prediction error which will be given in section 5.1.

The model selection methods considered here include AIC, BIC and ANOVA. For combining, the candidates are the 19 possible models in the three factor design, and the 167 possible models in four factor case (we include the null model as a candidate). Note that as usual, only hierarchical models are considered as candidate models. In three or four factor cases we consider in this work, it is computationally feasible to combine all possible models. 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.

Our goal in the empirical studies is not limited to a demonstration that ARM can work better than the alternative methods. Instead, the simulations and the examples are carefully chosen to help gain an insight into when model combining is advantageous relative to model selection in applications.

5.1 *Data Examples*

In this section, we are interested in finding out how the instability measures in section 3 can guide us to choose between combining and selecting. To that end, we will compare the performance of ARM with that of some model selection methods in the data sets used in section 3 on instability. For combining, we choose to combine all the possible hierarchical models with 3 or 4 factors.

For comparison of different methods, we take the loss to be the squared prediction error. 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 squared prediction error based on the difference between the test data and the estimated cell means.

We repeat the procedure 100 times to average out splitting variability. 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 squared prediction errors based on the 100 replications for each method are in Table 2. The standard errors of average squared prediction errors are given in the parentheses.

For the first four data sets, the advantage of ARM (relative to the other methods) increased as the instability of the data set increased. After the most stable data set by all measures, data set 2, ARM started to perform better than the model selection methods. Its advantage in risk reduction relative to the best of other methods increased from 2.6% to 15% and 18.8%. However, the advantage stopped increasing for the four factor models, holding steady at 12.5% and 12.3%. While the precise relationship between the instability

	ARM	ANOVA	AIC	BIC
Data set 1	13.464 (0.579)	15.761 (0.472)	15.761 (0.472)	15.761 (0.472)
Data set 2	2.216 (0.095)	1.731 (0.053)	2.384 (0.127)	2.384 (0.127)
Data set 3	53.164 (1.371)	59.125 (1.580)	54.573 (1.480)	56.974 (1.441)
Data set 4	32.501 (1.313)	-	40.027 (1.261)	41.537 (1.259)
Data set 5	0.035 (0.002)	-	0.040 (0.003)	0.042 (0.004)
Data set 6	13.295 (0.401)	-	15.166 (0.527)	15.457 (0.461)

Table 2

Comparing Model Selection to Combining by Squared Prediction Errors

measures and ARM performance is still to be worked out, the evidence here is consistent with our expectation that model combining is advantageous over model selection when the instability in model selection is high. So when a data set displays a high instability, it is probably necessary to consider model combining.

It is interesting to study what factors affect model instability. As we mentioned earlier, we expect the instability measures, and consequently the ARM performance to depend on sample size, the noise level and the model structure. In the following subsections, we will systematically investigate this relationship through simulations where we know the true model. Even though in reality we do not know the truth, this study will give us valuable insight into the property of selection and combining.

5.2 Simulations

The simulations start with the specification of a true model. The true cell means $\mu_{i_1 \dots i_\Phi}$ are calculated according to the model. In each cell J observations are generated from $N(\mu_{i_1 \dots i_\Phi}, \sigma^2)$ and 100 data sets are generated from the same true model. The loss is $\frac{1}{N} \sum_{i_1=1}^{I_1} \dots \sum_{i_\Phi=1}^{I_\Phi} (\mu_{i_1 \dots i_\Phi} - \hat{\mu}_{i_1 \dots i_\Phi})^2$. The average loss from these 100 replications is used as a Monte Carlo approximation of the risk of interest (average mean squared error). With each replication, the data are permuted 50 times to average out variability in splitting which occurs in model combining by ARM.

We consider several settings. Some fixed three and four factor models and some randomly generated models are analyzed. In all these settings, each factor has two levels.

Since we are primarily interested in cases where model selection may have difficulty in identifying the best model, in our simulations, we did not report

small σ^2 values. Our other simulation results showed that in some cases with small σ^2 values (as small as 0.1), ARM still had some advantage. The threshold σ^2 level at which ARM starts to outperform model selection remains an open question. This threshold level should depend on the true model structure as we will explore later.

5.2.1 Fixed Models

The simulation results are in Table 3. The standard errors of the risks are given in the parentheses.

Case 1 Three factors: Data are generated from $y = a + b + c + ab + \epsilon$ with $a_1 = 0.75, b_1 = 0.68, c_1 = 0.29, ab_{11} = 0.12$. Each cell has three replicates.

		ARM	ANOVA	AIC	BIC
Case 1	$\sigma^2 = 0.5$	0.076 (0.004)	0.080 (0.004)	0.069 (0.004)	0.069 (0.004)
	1.0	0.200 (0.011)	0.353 (0.032)	0.262 (0.016)	0.274 (0.019)
	1.5	0.389 (0.024)	0.940 (0.047)	0.662 (0.041)	0.730 (0.042)
Case 2	$\sigma^2 = 0.5$	0.106 (0.006)	0.079 (0.005)	0.072 (0.005)	0.074 (0.005)
	1.0	0.308 (0.013)	0.460 (0.036)	0.333 (0.019)	0.397 (0.026)
	1.5	0.515 (0.023)	1.044 (0.044)	0.799 (0.037)	0.906 (0.038)
Case 3	$\sigma^2 = 0.5$	0.090 (0.003)	0.087 (0.005)	0.073 (0.005)	0.066 (0.005)
	1.0	0.183 (0.009)	0.397(0.026)	0.271 (0.017)	0.297 (0.017)
	1.5	0.384 (0.025)	0.818 (0.038)	0.613 (0.034)	0.589 (0.030)
Case 4	$\sigma^2 = 0.5$	0.090 (0.001)	-	0.111 (0.001)	0.115 (0.001)
	1.0	0.148 (0.002)	-	0.174 (0.002)	0.199 (0.003)
	1.5	0.270 (0.041)	-	0.389 (0.060)	0.403 (0.062)

Table 3

Comparing ARM to Model Selection Methods with Three and Four Factor Models

From Table 3 we can see that as the noise level increased, the advantage of combining also increased. ARM showed no advantage at $\sigma^2 = 0.5$, and the reduction in risk by ARM over the best alternative was 23.7% and 41.2% in the other two noise levels respectively.

Case 2 Three factors: We kept the same model as in case 1 and the same parameter values except for one change: $ab_{11} = 0.50$. We increased the magnitude of the interaction term so it was not vague any more. As a result, we observed smaller advantage in ARM than in the previous case. ARM showed no advantage at $\sigma^2 = 0.5$, and the reduction in risk by ARM over the best

alternative was 7.5% and 35.5% in the other two noise levels respectively.

Case 3 Four factors: Data are generated from the model $y = a + b + c + d + \epsilon$, with $a_1 = 0.75, b_1 = 0.46, c_1 = 0.25, d_1 = 0.29$. Each cell has two replicates. The true model contained no obviously weak terms, but ARM still had an advantage even at $\sigma^2 = 0.5$. The reduction in risk by ARM was 16.7%, 32.5% and 34.8% respectively.

Case 4 The true model is the same as in Case 1. The difference is that the data come from an unbalanced design: each of the first 4 cells has 3 replicates and each of the remaining 4 cells has 4 replicates. In applying ARM, we randomly took 2 replicates from each cell as estimation data and used the remaining data for evaluation. This example demonstrated the advantage of ARM with unbalanced data. Since we can split the data cell by cell with different splitting proportions, ARM is easily adapted to unbalanced data.

The above four cases reaffirmed our analysis of model instability and ARM performance. Intuitively, higher noise level makes harder for model selection methods to identify the true model and that leads to higher instability and bigger ARM advantage. The simulation results are consistent with our expectation. However, we are more interested in finding out how the model structure affects model selection and ARM performance given a noise level. We expect to see bigger model instability and ARM advantage when the model contains some vague terms which add difficulty for selection methods. Comparison of Case 1 and Case 2 verified this relationship. We also expect increased model complexity would increase model selection instability and thus increase ARM advantage. The results of Case 3 are consistent with this expectation. Even though the true model does not contain vague terms, when it involves more factors and terms, ARM demonstrates bigger advantages. That makes intuitive sense as a more complicated model is usually more difficult to identify.

It is also of interest to compare the performance of AIC and BIC here. In our parametric setting, it is well known that BIC is consistent and AIC is not. However, for estimating the regression function (as is the focus in this work), BIC can perform much worse than AIC when some effects of the factors are substantially smaller than others (see, e.g. Yang, 2003). In our simulations, where sample sizes were small, AIC consistently outperformed BIC. We expect to observe better BIC performance with increased sample size.

5.2.2 *Random Models*

The above simulations suggest a relationship between model structure and the relative performance of model selection and ARM. In order to show that the above results hold in more general cases, we consider random models in this subsection.

We will consider three factor and four factor cases. In each case, we consider two settings where in one setting we purposely add weak terms in the true model.

A random model is first generated from the list of all possible models for three or four factors. Parameters for the main effects and the intercept are generated from uniform (0, 1). The parameters for the interaction terms are generated from uniform (0, 1) or uniform (0, 0.3). A noise level of $\sigma^2 = 1$ is used to generate the data.

Case 1 Three factors: All the parameters are assigned values from uniform (0, 1). In total 100 models are generated and 20 data sets are generated from each model. Each cell has three replicates. Each data set is permuted 50 times to smooth splitting variability in combining. The box plot in Figure 1 is based on the 100 risks from these 100 models.

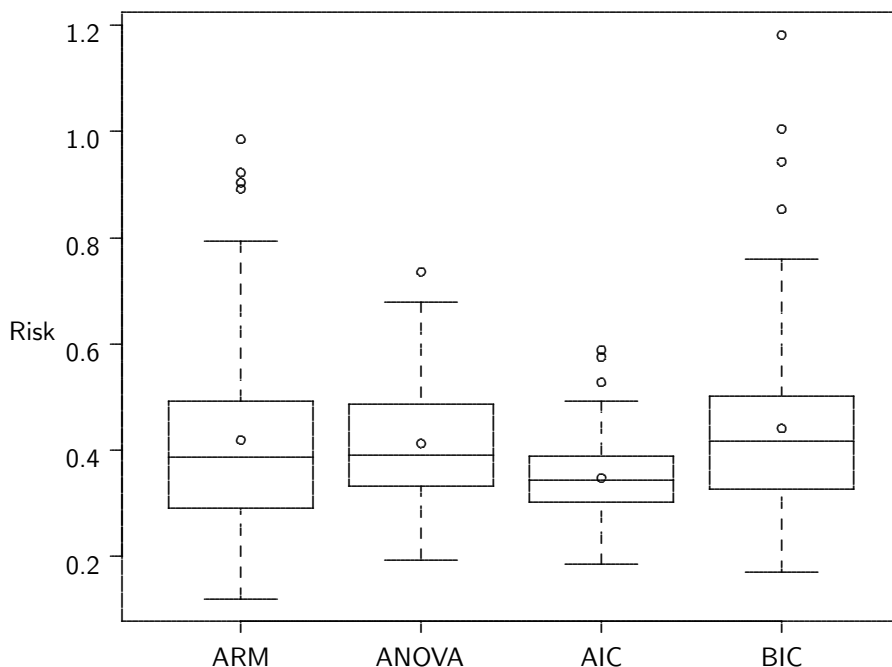


Fig. 1. Random Three Factor Models with Strong Interaction

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

The simulation results from random models are consistent with our earlier analysis: When the true model contained weak interaction terms (with parameters of interaction terms generated from uniform (0, 0.3)) which hampered the ability of model selection to identify the true model, the gain from ARM

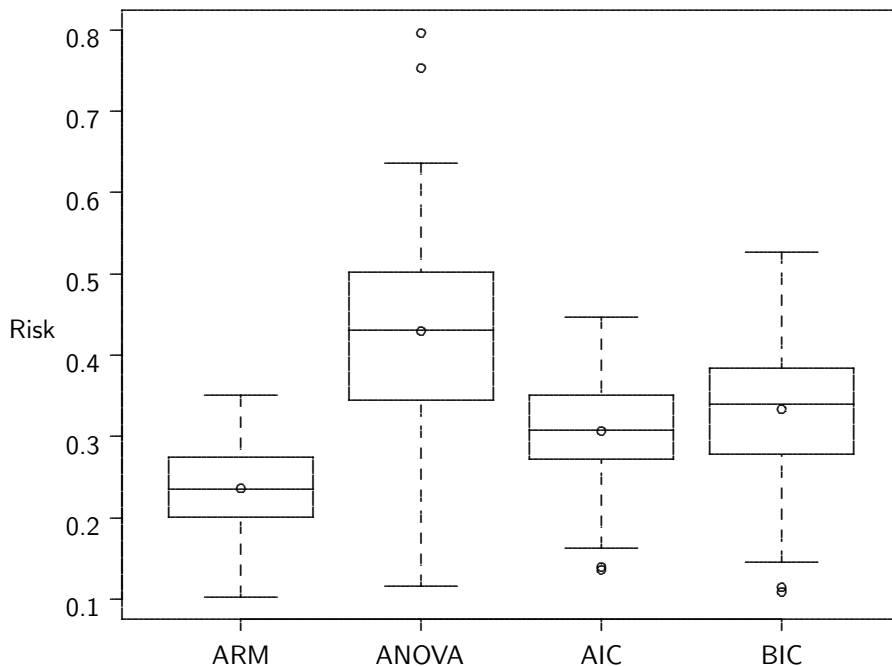


Fig. 2. Random Three Factor Models with Weak Interaction

Factors	Interaction	ARM	ANOVA	AIC	BIC
Three	Strong	0.419 (0.017)	0.413 (0.011)	0.348 (0.008)	0.441 (0.017)
Three	Weak	0.236 (0.005)	0.430 (0.012)	0.307 (0.006)	0.334 (0.008)
Four	Strong	0.500 (0.026)	0.475 (0.009)	0.435 (0.009)	0.597 (0.034)
Four	Weak	0.284 (0.006)	0.477 (0.010)	0.408 (0.010)	0.431 (0.013)

Table 4

Comparing ARM to Model Selection Methods with Random Models by Mean Risks

increased substantially. We also compared the two settings with four factor models and obtained similar results. The mean risks for the 100 random models in each case are in Table 4. The advantage of ARM also increased over the best alternative with an increased model complexity as we went from three factor to four factor case.

5.3 Which Factors Are More Important

As we have mentioned, given that the purpose of ANOVA is often to study factor effects, when model selection is appropriate, we prefer to choose model selection as it gives an answer directly. However, when the model selection

instability is high, the simple answer obtained through model selection is not trustworthy or even misleading.

In this case, a feasible approach is to take advantage of the more reliable estimates by ARM to answer the questions of interest. Intuitively, if we have accurately estimated cell means, that information may be helpful for understanding which factors are important. In particular, ARM's good performance in terms of prediction can be taken as a better representation of the data and we are less likely to err when we possess knowledge that is superior in the sense of more closely matching realizations of data outside the set at hand: by extension, we would hope the property would extend analogously from the experiment to the situation of interest which the experiment intended to produce information about. So in this subsection our goal is to investigate how to use ARM for assessing factor importance.

We start by averaging the cell means $\tilde{\boldsymbol{\mu}}$ over a factor. If the effect of the factor is small, after we average the cell means over that factor at each combination of the levels of the remaining factors, there should not be a big difference between the averaged cell means and the original cell means. Consequently, we can assess the importance of a factor by examining the magnitude of that difference between the averaged cell means $\tilde{\boldsymbol{\mu}}^*$ and the original reference cell means $\tilde{\boldsymbol{\mu}}$. If the difference is not big according to an appropriate criterion (e.g., practical significance), that factor may not be important.

One criterion for assessing the difference is the overall discrepancy across all the cells:

$$D_1 = \sqrt{\frac{\sum_{i_1, \dots, i_\Phi} (\tilde{\mu}_{i_1, \dots, i_\Phi} - \tilde{\mu}_{i_1, \dots, i_\Phi}^*)^2}{Ns^2}},$$

where N is the number of cells, and s^2 is the convex combination of the variance estimates from all the candidate models used in ARM, i.e., $s^2 = \sum \hat{W}_k \hat{\sigma}_k^2$. Here $\hat{\sigma}_k^2$ is the estimate of variance from model k , and \hat{W}_k is the weight assigned to that model in ARM.

Other reasonable criteria are the discrepancy in one cell of particular interest, e.g., the maximum discrepancy in one cell, i.e., $D_2 = \max_{i_1, \dots, i_\Phi} |\tilde{\mu}_{i_1, \dots, i_\Phi} - \tilde{\mu}_{i_1, \dots, i_\Phi}^*|/s$, or the discrepancy of the cell with the maximum ($\tilde{\mu}_{j_1, \dots, j_\Phi}$) or the minimum mean ($\tilde{\mu}_{k_1, \dots, k_\Phi}$) from ARM, i.e., $D_3 = |\tilde{\mu}_{j_1, \dots, j_\Phi} - \tilde{\mu}_{j_1, \dots, j_\Phi}^*|/s$, $D_4 = |\tilde{\mu}_{k_1, \dots, k_\Phi} - \tilde{\mu}_{k_1, \dots, k_\Phi}^*|/s$.

The above proposed criteria can also be used to evaluate the models chosen by model selection techniques (e.g. the ANOVA method) by assessing the fitted cell means from the selected model against the ARM $\tilde{\boldsymbol{\mu}}$.

We propose no criterion or cut off point with which to evaluate these measures and calculating their standard errors seems theoretically formidable. However, if the cell means estimated by ARM are reliable, the measures give an indication of the relative importance of the factors. The numbers can also be judged in terms of practical significance evaluated according to expert knowledge specific to a given problem.

We apply this procedure to the two four factor data sets we have analyzed. The discrepancy measures are summarized in Table 5. The factor over which the data table is collapsed is indicated in the parenthesis.

For data set 5, the measures tend to suggest that the 4 factors are similarly important; but for data set 6, from the perspective of D_1 and D_3 , factor B is substantially less important than the other factors.

	D1	D2	D3	D4
Data set 5	0.726(A)	1.104(A)	0.905(A)	1.091(A)
	0.543(B)	0.770(B)	0.770(B)	0.733(B)
	0.726(C)	1.091(C)	0.780(C)	1.091(C)
	0.579(D)	0.979(D)	0.979(D)	0.444(D)
Data set 6	0.704(A)	1.283(A)	0.983(A)	0.470(A)
	0.255(B)	1.712(B)	0.312(B)	0.271(B)
	0.630(C)	0.790(C)	0.463(C)	0.764(C)
	0.812(D)	1.134(D)	1.132(D)	0.321(D)

Table 5
Analyzing Factor Effects of Data Set 5 and 6

Applying the ANOVA method to the original data concludes that all 4 factors are significant in both data sets. For data set 5, the overall discrepancy measure between the combined cell means $\hat{\mu}$ and the fitted cell means from the ANOVA model $\tilde{\mu}$ is $D_1 = 0.374$. The maximum discrepancy is $D_2 = 0.471$. The discrepancy in the cell with the maximum or minimum cell mean is $D_3 = 0.471$, $D_4 = 0.184$ respectively. For data set 6, the corresponding measures are: $D_1 = 0.258$, $D_2 = 0.568$, $D_3 = 0.099$, $D_4 = 0.073$. For the two data sets, the D_2 values are not small, indicating that the estimates based on the ANOVA method are significantly different from the ARM estimates. Observing that D_3 and D_4 are all very small for data set 6, if our goal is to find the cell that maximizes or minimizes the mean response, it is perhaps notable that there is no disagreement between ARM and the ANOVA method in this case.

Next we use one simulation to show the potential advantage of the method we propose in assessing the importance of factors. Consider a 2^3 design with

the true model $y = a + b + c + \epsilon$, with $a_1 = 0.80$, $b_1 = 0.50$, $c_1 = 0.10$. Choose $\sigma^2 = 0.25$. Generate 100 data sets from this model with three replicates in each cell. For each data set, compute the cell means estimate $\tilde{\mu}$ by ARM. Consider the overall discrepancy measure D_1 . In all the 100 data sets, we found the ranking $D_1(A) > D_1(B) > D_1(C)$ held, where the letter indicates the factor over which the data table is collapsed. The means of the three measures were 5.43, 3.13 and 1.87 respectively. In this example, the discrepancy measure was rather reliable in ranking the importance of the factors, and thus gave a relative answer to the question of importance of factors. On the other hand, over 30% of the times, the model selected by ANOVA failed to rank the importance of the factors correctly.

6 Summary and Conclusion

While model averaging solves some problems, it is not automatically better than model selection. In addition to being hard to interpret, model combining can perform worse than model selection in terms of estimation risk. Understanding when combining has an advantage over selection would be a step forward in research on model combining. It would also be helpful to systematically compare model combining and model selection methods beyond a few selected examples (as was typically done in previous publications). In this research on combining factorial models, we have worked along these lines.

Based on the studies we have done so far on the property of ARM and model selection, we have found:

- ARM and model selection performance is related to model selection instability. When a model selection criterion has no difficulty in identifying the best model for a given data set, model selection usually outperforms model combining.
- ARM has a substantial advantage over model selection when there is high uncertainty in model selection. In simulations, when the true models contain weak interaction terms or more factors, ARM usually has a bigger advantage.
- Bootstrap, perturbation, and sequential instabilities properly measure uncertainty in model selection from somewhat different angles. The data examples showed a monotonic association between the instability measures and the relative performance of ARM compared to model selection methods. We recommend the use of these instability measures in factorial data analysis.

In the cases examined in this paper, the various model selection techniques vie with each other in terms of risk while ARM substantially outperforms model

selection when model selection instability is not negligible. This result is in keeping with theoretical work on ARM. Based on our experience analyzing textbook factorial data, model selection instability is usually high with 4 or more factors and up to 3 replications (a configuration commonly seen in industrial applications). Thus, ARM is an attractive alternative to model selection in the ANOVA setting. We have here demonstrated its strength when prediction (or estimating the means) is important. But results are also encouraging for the possibility of developing a method for evaluating the importance of factors.

The positive results in this investigation suggest future work in which the relationship between instability and ARM performance can be elaborated in the hope of developing a method for determining more precisely when ARM is likely to have an advantage in terms of risk over model selection. Further work is also warranted in developing an ARM based method for evaluating factor importance, or more broadly for assessing the consequences of making decisions based on factorial data, including constructing ARM based confidence intervals.

This paper is limited to implementing ARM in the ANOVA setting and understanding its performance there as an alternative to model selection. Thus, comparisons are made to common model selection methods with well known properties to provide a clear benchmark. Comparison of ARM to other model combining techniques is also appropriate. Bayesian Model Averaging (BMA) and other model combining techniques could be extended to this setting. Kass and Raftery (1995) have reported an acceptable large sample approximation for Bayes factors based on BIC. Using this approximation to calculate BMA weights has the advantage of being less computationally intensive than one involving exact evaluation of the posterior probabilities of candidate models. AIC can be used instead of BIC to yield yet another weighting scheme described by Buckland et al. (1997). Chen and Giannakouros (2006) have considered such approaches in subsequent work in other settings and have started preliminary explorations in the ANOVA setting.

We have written a series of R functions to implement ARM for ANOVA models. The software can be obtained at on the World Wide Web.

Appendix: Theory and Proof

Consider the setup in sections 2 and 4.

A risk bound for ARM

For a vector $\mathbf{a} = (a_1, \dots, a_w)$, let $\|\mathbf{a}\|_\infty = \max_{1 \leq i \leq w} |a_i|$. Let $\hat{\boldsymbol{\mu}}^{(k)}$ be the estimate of $\boldsymbol{\mu}$ and $\hat{\sigma}_k^2$ be the estimate of σ^2 from model k based on the first part of the data.

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

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

Condition 2: There exist constants $0 < \xi_1 \leq 1 \leq \xi_2 < \infty$ such that

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

with probability one for all $1 \leq k \leq K$.

The above conditions are satisfied if all the cell means and the error variance are upper and lower bounded (away from 0 for the variance case) 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 (2001), for the theoretical result, we study a slightly different estimator from that given in Section 4. Recall that the data are randomly split into two parts with J_1 and J_2 observations in each cell, with the first part of the data $Z^{(1)}$ containing $n_1 = J_1 N$ observations and the second part of the data $Z^{(2)}$ containing $n_2 = J_2 N$ observations in total. Let $n = n_1 + n_2$. Stack $Z^{(2)}$ into one vector $\mathbf{Y} = (y_{n_1+1}, y_{n_1+2}, \dots, y_n)$ 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 m_i . Note that if y_i and y_j are in the same cell, we have $m_i = m_j$. Let $\hat{m}_i^{(k)}$ be the estimate for m_i from model k based on $Z^{(1)}$.

For $i = n_1 + 1$, let $W_{k,i} = 1/K$ and for $n_1 + 2 \leq i \leq n$, let

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

Let

$$\tilde{m}_i = \sum_{k=1}^K W_{k,i} \widehat{m}_i^{(k)}. \quad (1)$$

Note that according to which cell m_i belongs to, $\{\tilde{m}_i\}$ can be written as $\{\tilde{m}_{r,j}\}$ where $r = 1, \dots, N$ denotes the cells and $j = 1, \dots, J_2$ denotes the replicates. Define

$$\tilde{\mu}_r = \frac{1}{J_2} \sum_{j=1}^{J_2} \tilde{m}_{r,j}.$$

For two vectors $\mathbf{a} = (a_1, \dots, a_p)$, let $\|\mathbf{a}\|^2 = \frac{1}{p} \sum_{i=1}^p a_i^2$.

Theorem 1: Assume that the errors are Gaussian and that Conditions 1 and 2 are satisfied. Then the average mean squared error of the combined estimator satisfies

$$E \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq (1 + \xi_2 + 9\tau/2) \inf_{k \geq 1} \left(\frac{2\sigma^2 \log K}{n_2} + \frac{1}{\xi_1} E \|\hat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} E(\hat{\sigma}_k^2 - \sigma^2)^2 \right),$$

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

This theorem indicates the risk for the combined estimator stays close to $\inf_k E \|\hat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|^2$, the minimum risk for estimating $\boldsymbol{\mu}$ among the competing estimators. Of course, the minimum risk is not directly achievable because we do not know which model is true. From the above theorem, up to a constant factor and an additive penalty $(\log K)/n$, the combined procedure achieves this best performance plus the risk of variance estimation. We emphasize that the risk bound in the theorem holds for each sample size. This is in sharp contrast to an asymptotic expression on the risk, which can be misleading when there is much uncertainty in model selection because the asymptotic behavior is typically unreliable for such a situation. Note that when ξ_1 and ξ_2 are around 1 and when τ is not large, the multiplicative factor is good. Roughly speaking, if when the sample size n increases, the estimators chosen to be combined are more and more accurate so that $\tau \rightarrow 0$ and ξ_1 and ξ_2 converge to 1, the multiplicative factor approaches 2.

Note that the estimator $\tilde{\boldsymbol{\mu}}$ in Theorem 1 as defined by (1) is not exactly the same as $\tilde{\boldsymbol{\mu}}$ given in Section 4. The modified estimator here is slightly more complicated and computationally more costly (but with the theoretical bound). As in Yang (2001), the simpler estimator is recommended in practice.

Proof of Theorem 1:

Proof: For simplicity, denote $\mathbf{m} = \{m_i\}$, $\widehat{\mathbf{m}} = \{\widehat{m}_i\}$, $\widetilde{\mathbf{m}} = \{\widetilde{m}_i\}$, where \widetilde{m}_i is the combined estimate of the mean of the cell where y_i belongs to for $i = n_1 + 1, \dots, n$. Let

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

and

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

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

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

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

$$E_{n_1} \left(\log \frac{\widehat{\sigma}_{k^*}^2}{\sigma^2} + \frac{(y_i - \widehat{m}_i^{(k^*)})^2}{\widehat{\sigma}_{k^*}^2} - \frac{(y_i - m_i)^2}{\sigma^2} \right) = \frac{(\widehat{m}_i^{(k^*)} - m_i)^2}{\widehat{\sigma}_{k^*}^2} + \frac{\sigma^2}{\widehat{\sigma}_{k^*}^2} - 1 - \log \frac{\sigma^2}{\widehat{\sigma}_{k^*}^2}. \quad (3)$$

Observe that q^{n_2} can be rewritten as

$$\begin{aligned} &\frac{1}{K} \sum_{k=1}^K \frac{1}{\sqrt{2\pi\widehat{\sigma}_k^2}} \exp\left(-\frac{1}{2\widehat{\sigma}_k^2}(y_{n_1+1} - \widehat{m}_{n_1+1}^{(k)})^2\right) \\ &\times \frac{\frac{1}{K} \sum_{k=1}^K \frac{1}{(\sqrt{2\pi\widehat{\sigma}_k^2})^2} \exp\left(-\frac{1}{2\widehat{\sigma}_k^2}((y_{n_1+1} - \widehat{m}_{n_1+1}^{(k)})^2 + (y_{n_1+2} - \widehat{m}_{n_1+2}^{(k)})^2)\right)}{\frac{1}{K} \sum_{k=1}^K \frac{1}{\sqrt{2\pi\widehat{\sigma}_k^2}} \exp\left(-\frac{1}{2\widehat{\sigma}_k^2}(y_{n_1+1} - \widehat{m}_{n_1+2}^{(k)})^2\right)} \end{aligned}$$

$$\times \dots \times \frac{\frac{1}{K} \sum_{k=1}^K \frac{1}{(\sqrt{2\pi\hat{\sigma}_k^2})^{n_2}} \exp\left(-\sum_{i=n_1+1}^n \frac{1}{2\hat{\sigma}_k^2} (y_i - \hat{m}_i^{(k)})^2\right)}{\frac{1}{K} \sum_{k=1}^K \frac{1}{(\sqrt{2\pi\hat{\sigma}_k^2})^{n_2-1}} \exp\left(-\sum_{i=n_1+1}^{n-1} \frac{1}{2\hat{\sigma}_k^2} (y_i - \hat{m}_i^{(k)})^2\right)}.$$

Let $p_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - m_i)^2}{2\sigma^2}\right)$ and $g_i = \sum_{k=1}^K W_{k,i} \frac{1}{\sqrt{2\pi\hat{\sigma}_k^2}} \exp\left(-\frac{(y_i - \hat{m}_i^{(k)})^2}{2\hat{\sigma}_k^2}\right)$ for $n_1 + 1 \leq i \leq n$. By the construction of $W_{k,i}$, $\log(p^{n_2}/q^{n_2}) = \sum_{i=n_1+1}^n \log\left(\frac{p_i}{g_i}\right)$. Together with (2) and (3), we have

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

From the familiar relationship between the Kullback-Leibler divergence and the squared Hellinger distance, we have

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

We next lower bound the Hellinger distance. Let p and g be two probability densities on the real line with respect to a measure ν , with means m_p and m_g , variances $0 < \sigma_p^2 < \infty$ and $0 < \sigma_g^2 < \infty$ respectively. Then from Lemma 1 of Yang (2004),

$$\int (\sqrt{p} - \sqrt{g})^2 d\nu \geq \frac{(m_p - m_g)^2}{2(\sigma_p^2 + \sigma_g^2) + (m_p - m_g)^2}.$$

Under Conditions 1 and 2, it is easy 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 $\tilde{m}_i = \sum_{k=1}^K W_{k,i} \hat{m}_i^{(k)}$, we have

$$E_{n_1} \log\left(\frac{p_i}{g_i}\right) \geq \frac{(\tilde{m}_i - m_i)^2}{\sigma^2(2(1 + \xi_2) + 9\tau)}.$$

Together with (4), we have

$$\sum_{i=n_1+1}^n E \left(\frac{(\tilde{m}_i - m_i)^2}{\sigma^2(2(1 + \xi_2) + 9\tau)} \right) \leq \log K + \frac{n_2}{2} E \left(\frac{\|\hat{\mathbf{m}}^{(k)} - \mathbf{m}\|^2}{\hat{\sigma}_{k^*}^2} + \frac{\sigma^2}{\hat{\sigma}_{k^*}^2} - 1 - \log \frac{\sigma^2}{\hat{\sigma}_{k^*}^2} \right).$$

Recall that $\{\widetilde{m}_i\}, \{m_i\}$ can be written as $\{\widetilde{m}_{r,j}\}, \{m_{r,j}\}$ where $r = 1, \dots, N$ and $j = 1, \dots, J_2$. By convexity of the square function, we have

$$E \left(\left(\frac{1}{J_2} \sum_{j=1}^{J_2} \widetilde{m}_{r,j} \right) - m_{r,j} \right)^2 \leq \frac{1}{J_2} \sum_{j=1}^{J_2} E (\widetilde{m}_{r,j} - m_{r,j})^2.$$

Note that $\frac{1}{J_2} \sum_{j=1}^{J_2} \widetilde{m}_{r,j} = \widetilde{\mu}_r$ and $\{m_{r,j}\}$ are same for the same r , i.e., $m_{r,j} = \mu_r$. Also replace $\|\widehat{\mathbf{m}}^{(k)} - \mathbf{m}\|^2$ with $\|\widehat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|^2$, we have

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

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

$$E \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq (1 + \xi_2 + 9\tau/2) \inf_{k \geq 1} \left(\frac{2\sigma^2 \log K}{n_2} + \frac{1}{\xi_1} E \|\widehat{\boldsymbol{\mu}}^{(k)} - \boldsymbol{\mu}\|^2 + \frac{C(\xi_1, \xi_2)}{\sigma^2} E(\widehat{\sigma}_k^2 - \sigma^2)^2 \right),$$

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

This completes the proof of Theorem 1.

Acknowledgments This research was supported by the United States National Science Foundation CAREER Award Grant DMS0094323. We thank the editors and two referees for many helpful comments.

References

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. pp. 267–281.
- Breiman, L., 1996. Heuristics of instability and stabilization in model selection. *Annals of Statistics* 24, 2350–2383.
- Buckland, S. T., Burnham, K. P., Augustin, N. H., 1997. Model selection: An integral part of inference. *Biometrics* 53, 603–618.
- Chen, L., Giannakouros, P., August 2006. Model combining in survival analysis, paper presented at the annual Joint Statistical Meetings, Seattle, WA.
- Diaconis, P., Efron, B., 1983. Computer-intensive methods in statistics. *Scientific American* 248, 116–130.

- Garcia-Diaz, A., Phillips, D. T., 1995. Principles of Experimental Design and Analysis. Chapman & Hall, London, UK.
- Hoeting, J. A., Madigan, D., Raftery, A., Volinsky, C. T., 1999. Bayesian model averaging: a tutorial with discussion. *Statistical Science* 14, 382–417.
- Kass, R. E., Raftery, A., 1995. Bayes factors. *Journal of the American Statistical Association* 90, 773–795.
- McClean, R. A., Anderson, V. L., 1984. Applied Factorial and Fractional Design. Marcel Dekker, Inc. , NY.
- Montgomery, D. C., 1997. Design and Analysis of Experiments. John Wiley & Sons, Inc. , NY.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W., 1996. Applied Linear Statistical Models. The McGraw-Hill Companies, Inc. , Boston.
- Raftery, A., 1996. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* 83, 251–266.
- Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Vardeman, S. B., Jobe, J. M., 2001. Basic Engineering Data Collection and Analysis. Duxbury Press, North Scituate, MA.
- Yang, Y., 2001. Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–588.
- Yang, Y., 2003. Can the strengths of AIC and BIC be shared? *Biometrika* 92, 937–950.
- Yang, Y., 2004. Combining forecasting procedures: some theoretical results. *Econometric Theory* 20, 176–222.