Journal of Machine Learning Research 4 (2003) 683-712 Submitted 10/02; Revised 8/03; Published 10/03

Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored

Bertrand Clarke BERTRAND@STAT.UBC.CA
Department of Statistics
University of British Columbia
Vancouver, BC V6T 1Z2, Canada

Editors: Bin Yu

Abstract

We compare Bayes Model Averaging, BMA, to a non-Bayes form of model averaging called stacking. In stacking, the weights are no longer posterior probabilities of models; they are obtained by a technique based on cross-validation. When the correct data generating model (DGM) is on the list of models under consideration BMA is never worse than stacking and often is demonstrably better, provided that the noise level is of order commensurate with the coefficients and explanatory variables. Here, however, we focus on the case that the correct DGM is not on the model list and may not be well approximated by the elements on the model list.

We give a sequence of computed examples by choosing model lists and DGM's to contrast the risk performance of stacking and BMA. In the first examples, the model lists are chosen to reflect geometric principles that should give good performance. In these cases, stacking typically outperforms BMA, sometimes by a wide margin. In the second set of examples we examine how stacking and BMA perform when the model list includes all subsets of a set of potential predictors. When we standardize the size of terms and coefficients in this setting, we find that BMA outperforms stacking when the deviant terms in the DGM 'point' in directions accommodated by the model list but that when the deviant term points outside the model list stacking seems to do better.

Overall, our results suggest the stacking has better robustness properties than BMA in the most important settings.

Keywords: Key words: Bayes model averaging, stacking, robustness, model selection.

1. Introduction and an Example

Consider the following toy problem. Suppose the true model, i.e., the data generating model (DGM), that produced the data we want to analyze, is a linear regression model with outcomes Y, IID $N(0,\sigma^2)$ errors denoted ε, 12 explanatory variables $X_0,...,X_{11}$, and corresponding parameter vector $\beta = (\beta_0,...,\beta_{11})$. We take $X_0 = 1$ to be the constant term. Thus we have a supermodel $Y = X\beta + \varepsilon$. Suppose that the investigators do not know what the true model is but have identified 3 models that they think are plausible, and each of them is a submodel of the true model. The 3 models are $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, $Y = \beta_0 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$ and $Y = \beta_0 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon$. Together, these three models form a set we call the model list $\mathcal{M}$. Since they are disjoint in the sense of having no explanatory variables in common, we regard them as forming the vertices of a triangle in some model space. See Figure 1 in which we have represented models by the indices of their explanatory variables.

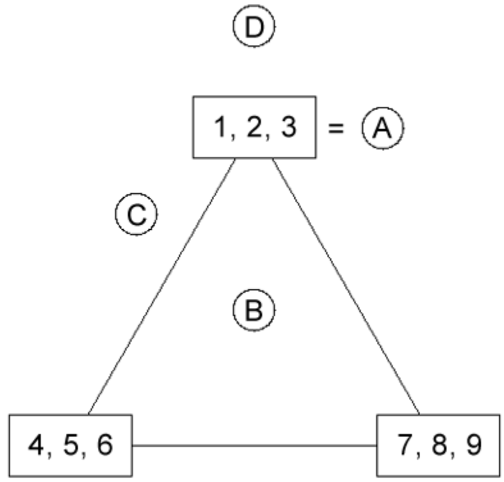

Figure 1: The Triangle Example.

Next, let us consider four scenarios differing in which model is actually true. Scenario A has $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ as the DGM so the true model is on the model list. Scenario B has $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_7 X_7 + \varepsilon$ as the DGM; it is not on the model list but has elements from all three models on the list. Scenario C has $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_{10} X_{10} + \varepsilon$ as the DGM; it is not on the model list. Without the X_{10} term it would be a combination of the first two models on the model list. The extra variate, X_{10} means that it deviates from being a combination of the first two models. Finally, consider scenario D in which $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{10} X_{10} + \beta_{11} X_{11} + \varepsilon$ is the DGM. Clearly, it is closest to the first model (which is exactly correct in scenario A) but further by an equal amount from the other two models on the list.

To proceed, one must have a way to use the model list to get at the true model. For comparison purposes, consider three techniques. One is called stacking, (see work by Wolpert, 1992, Breiman, 1996a, and Smyth and Wolpert, 1998). It finds coefficients based on a technique like cross-validation as described in Section 2.1 below. The other is Bayes model averaging (BMA), which assigns weights on the basis of posterior probabilities for the models on the list. The third will be to choose the model having the largest posterior probability, essentially the Bayes information criterion, BIC.

In the supermodel, β has 11 entries; the submodels are formed by setting some of the entries β_i in β equal to zero. Now, if we generate β's from IID $N(0,1)$'s and then generate an $n \times 11$ matrix X using $N(0,1)$'s as well, we can get an $n \times 1$ vector Y from $N(X\beta, 1)$. This means we can get an estimate $\hat{\beta}$, taken here to be the posterior mean, for each model in the model space. Next, we combine the models with their β's by averaging, either by stacking or BMA to get an overall fitted value $\hat{Y}$. For both cases we need coefficients, $\alpha_{i,S}$ for stacking and $\alpha_{i,B}$ for BMA with $i = 1,2,3$, to serve as weights for the 3 models in the two averaging schemes. The $\hat{\beta}$'s that result from the model average then are $\hat{\beta}_S = \sum \alpha_{i,S} \hat{\beta}_i$ and $\hat{\beta}_B = \sum \alpha_{i,B} \hat{\beta}_i$. (The fitted values Y_S and Y_B are obtained from submatrices of X with nonzero entries corresponding to the explanatory variables in the models.)

Here, we evaluate how good a model is by its risk, $R = E||\beta - \hat{\beta}||^2$, where $\hat{\beta}$ is formed from adding the posterior means of the models using weights from stacking or BMA. This risk admits a predictive interpretation that is in the same spirit as Hoeting et al (1999, Section 6) who used a cross-validation criterion. Moreover, evaluating R is intuitively equivalent to examining the regression functions directly – R cannot be small unless the model average nearly coincides with the true model. We get an estimate of R by iterating the procedure described above, varying the model taken as true (A, B, C, or D) and the form of averaging used. That is, we are fixing the model list and varying the DGM to see how well the list performs.

If we choose a sample size of 100, we get a table of risks for 12 cases, depending on the DGM and the technique of model averaging. Here, we generated the Y_i's from $N(X\beta, \sigma^2)$ with $\beta = \underline{1}$, a vector with all entries one, and $\sigma^2 = 1$. The results follow.

TRUE MODEL	BMA	STACKING	CHOICE
A	0.03	0.043	0.03
B	0.56	0.45	0.60
C	0.56	0.38	0.57
D	0.54	0.57	0.54

Table 1: Risks for four models and three techniques

There are three things to note about Table 1. First, the third technique, model choice, in which one chooses a fixed model to use never has the lowest risk for any row, except when it ties with BMA. This is no surprise: Advocates of model averaging in general argue that averaging models outperforms any specific model choice technique.

Indeed, in a series of papers, Yang (2003a, 2003b, 2003c) contrasts model selection and model aggregation. This is done with several techniques, including AIC, BIC and BMA, in several contexts, time series and regression, from the standpoint of risk, estimation accuracy, and predictive accuracy. The overall import of this work is that combined forecasting procedures typically have risks with rates of convergence that are the same as if the best forecasting procedures were known. Also, selection tends to work well relative to mixing only when the noise is small but that as the noise increases the performance of mixing techniques improves more and more over selection, under a risk criterion.

Second, the risks for the BMA when the true model is B, C, or D are very close, whereas those for stacking are different. This suggests that BMA is less sensitive to the geometry of the model list than stacking. Indeed, we suggest BMA is sensitive primarily to how close the DGM is to the closest members of the model list (see Berk, 1966). Third, stacking had the lowest risk in B and C where the true model was not on the model list but had some terms in common with models on the list. In D, BMA won by a little, but it is seen that the extra terms in D not in $\mathcal{M}$ amount to a larger noise rather than an explanatory variable. So, BMA is winning because it deals with model uncertainty optimally not because it deals well with bias.

Note that in Table 1 we used a fixed β instead of random β so R is not strictly a Bayes risk. However, this is one of the simplest generic cases we could think of. The explanatory variables are uncorrelated and as few as possible, consistent with having interesting cases; there is no variability in the coefficients which have a single typical value; the variability in the noise, σ, is the same as the variability in the explanatory variables. Thus, we can attribute differences in performance primarily to the location of the DGM relative to the model list. Henceforth, we include variability in the coefficients β so that we really are examining a Bayes risk.

As a practical matter, in the cases we examined, the qualitative behavior of risks and Bayes risks was the same provided the extra variability introduced by use of a prior within regression models was commensurate with the variability in the noise term and explanatory variables. If one permits correlation among the explanatory variables – as one would expect in practice – R remains a meaningful quantity to compute, however its predictive interpretation and its usefulness as an evaluation of the regression estimator is weakened.

One way to express the difference between two models is to count the number of terms by which they differ. That is, given two models M_1 and M_2 add the number of terms which are in M_1 but not M_2 to the number of terms in M_2 but not M_1. This is the cardinality of the symmetric difference applied to the applied to the terms in the model. (In Section 5, we argue this is reasonable in regression settings with uncorrelated explanatory variables; much beyond this context it may be a serious oversimplification.) If we represented the model with X_1, X_2 and X_3 as explanatory variables by 123 (and the other models similarly) then, for instance, in scenario A, the DGM is on the model list so the symmetric difference on the terms between the DGM and 123 is zero. However, in scenario B the symmetric difference between the DGM and 123 is three since the DGM has two terms (X_4 and X_7) that aren't in 123 and one term X_3 is not in in the DGM.

If we try to relate the better form of averaging to the sum of the symmetric differences between the DGM and all members of the model list we get Table 2. It is seen that BMA seems to win near and far from the DGM, i.e., for DGM's A and D, but stacking wins on the midrange distances. Thus, Bayes optimality appears to drop off rapidly as the DGM deviates from the model list. However, BMA may recover when the distance from the DGM to the model list can be interpreted as noise rather than bias, as in D. This may occur because BMA depends on the likelihoods more than stacking does while stacking is more data driven than BMA is.

		Distance to:			
DGM	WINNER	123	456	789	SUM
A	BMA	0	6	6	12
B	Stacking	3	5	5	13
C	Stacking	3	5	7	15
D	BMA	3	7	7	17

Table 2: Distances in model space

It is important to distinguish between model averaging and model combination (e.g., Minka, 2000). Model averaging assumes we have several models, as in the triangle example and we form a weighted sum of them. By contrast, model combination is the 'all subsets' case in which we would find coefficients for each term we were willing to include. Usually, this is done by weighting all submodels of one supermodel formed from all the terms in the models one wishes to combine. Breiman (1996a) claimed that the biggest gains arise when sets of dissimilar functions are used. If this claim is substantially true then the distinction between averaging and combination is mostly a function of the metric geometry of the model lists rather than the functional forms.

An approach intermediate between the two extremes of using many individual terms or many dissimilar functions is adaptive. For instance, one can start with an exhaustive list and prune out inappropriate models. In effect, one is using the data to choose a list of models to average over. Otherwise put, one can choose a big model and use appropriate regularizers. One way to do this is via the LASSO, Tibshirani (1994). This method sets various coefficients to zero by a shrinkage criterion. Kam (2002) uses a different technique in a neural nets context to eliminate elements of a general model from further consideration.

A key point here is to regard model uncertainty as relative to a model list, which we take to be a finitely parameterized subset of a model space. A model space is a large, presumably non-parametric collection of models in which it is safe to assume the DGM lies. Posterior probabilities are one well known description of model uncertainty, relative to a model list. Here, we also consider approximation error, namely, how well the true DGM is approximated by elements of the model list. To see how this affects our inferences, we examine settings in which the DGM is not readily expressed in terms of the model list.

An idealistic Bayesian would argue that one should use BMA (or whatever other procedure emerged from a Bayes risk optimization). In this case, the Bayesian would want to put a prior on the full model space, or a countable dense subset, so that approximating the DGM would be assured. Admitting the difficulty of this, the orthodox Bayesian would want to choose a finite model list to generate an approximation to the DGM and verify that the loss due to that approximation is small. This approach is difficult too, although philosophically consistent. The main problems would arise from trying to ensure the posterior probabilities of the models on the model list converge properly. Essentially this is the view taken by Hoeting et al. (1999).

Draper (1995) has a more conceptually satisfying framework. He distinguishes between a 'within structure variance' (WSV) and a 'between structure variance' (BSV). In BMA, the WSV is the weighted variance over the models one is averaging and the BSV is often neglected partially because it is difficult to define classes of structures and give meaningful distributions for them. Nevertheless, the intuition behind the idea can be implemented. One example is work by Gustafson and Clarke (2003). Although we do not examine BSV directly here, we argue the concerns it represents are encapsulated by our focus on the degree of model mis-specification weighting methods can tolerate.

Unfortunately, stacking doesn't have an associated treatment for the WSV like BMA does. Consequently, we have used a risk criterion which can be regarded, in some cases, like a cumulative average prediction error. This can be evaluated for both BMA and stacking providing a common performance standard. For comparison purposes this is better because, as a criterion, prediction permits all methods to compete equally for accuracy. We would not, in general, want to restrict our attention to methods which had a well defined treatment for WSV when we knew other methods gave better predictive performance.

The overall view developed here, from the triangle and other examples presented later, is primarily for the regression case where the sources of variability in parameters, noise terms, and uncorrelated explanatory variables is roughly comparable and the sample size is moderate. In this context, moderate means large enough to permit discernment among models but small enough that uncertainty remains. First, when the DGM is on the model list, or very close to an element on the model list BMA wins over stacking. Second, on the closed convex hull of the model list outside the BMA domain some form of stacking wins over BMA. In general, for models of the form $Y = f(X) + \varepsilon$, where f is in a class of functions, X is an explanatory variable and ε is a noise term, the convex hull consists of all functions of the form $\alpha f_1(X_1) + (1-\alpha) f_2(X_2)$ for $0 \leq \alpha \leq 1$. Here, our f's are linear in X with coefficients β so the model lists are finite dimensional so the convex hull is closed already, in the sense that any convergent sequence converges to a point in it. Alternatively, one can imagine the closed convex hull formed from the mixtures from each parametric model because the mixture is a good summary for all members of a parametric model. Third, for deviations of the DGM beyond the closed convex hull, it is more typical for stacking to win than for BMA to win, although this is in terms of magnitude of deviations. For many smooth and low variability directions of deviation of the DGM from the model list for which BMA wins over stacking. However, stacking usually wins over BMA for deviations of the DGM from the model list that involve nonsmooth or other high variability functions. As one includes correlations among stochastic quantities or permits perturbations with size or variability outside the range the model list accommodates, this interpretation becomes less tenable.

This amounts to a variance/bias tradeoff: When all random quantities in a regression setting are of roughly comparable variability, BMA tends to win over stacking when bias is relatively low but stacking tends to win over BMA when bias is relatively high. That is, the envelope around the closure of the model list on which BMA approximates the DGM well and outside of which stacking will typically win, will depend on the relative sizes of variances of the noise term, the coefficients, and the explanatory variables and how these compare to the approximation error.

As noted above, our results suggest that direction of deviation matters as well as magnitude: BMA usually loses out to stacking for unfortunate directions of deviation from the DGM to its best approximation using the model list. At this time we are unable to formulate the notion of an 'unfortunate direction' any better than to argue heuristically: Regard the region where BMA wins over stacking as a manifold in the model space. At each point on this manifold there is a finite dimensional tangent space containing all the directions in which one can perturb the point while remaining inside the manifold. (That is, there must be a smooth curve in the manifold which passes through the point and has that direction as its tangent vector at the point.) An unfortunate direction is a perturbation that takes the point outside the manifold.

As a practical matter, identifying these unfortunate directions in general is tentative. In the computational work reported here, unfortunate directions correspond to deviation terms in the DGM, biases, that are sufficiently difficult to approximate by elements of the model list. This includes, but is not limited to, products of a polynomial with the indicator function of a set, and functions that have a term of much higher or lower variability than the other terms. This latter category includes, for instance, polynomials of degree higher by, say, two or more, than those in the model list. It also includes polynomials with a slightly different exponent, say 1/3 or more, when the exponents are less than one. In addition, unfortunate directions includes cases where the deviation is smooth and has variability similar to those on the list, but is different in functional form. This includes cases where explanatory variables, or function of them are missing. It appears that missing two or more 'reasonable' variables impairs the approximative power of the model list enough to make BMA suboptimal. The results described here are broadly consistent with the suggestion that unfortunate directions of deviation are functions $f(x)$ in which f has a higher rate of change with increasing x than the functions on the model list.

In the absence of such unfortunate directions, BMA is best. However, the sensitivity of BMA to the direction of perturbation may be so great that even a small magnitude perturbation, well within experimental or sampling error, may make BMA underperform relative to stacking. Indeed, this viewpoint is consistent with the common practice of using BMA more for variable selection than for obtaining a regression function, see Hoeting et al. (1999, Section 4.1). Our effort here is to use model averaging as a regression technique in which we imagine the true regression function as situated in an infinite dimensional (nonparametric) function space. For this reason, we have great latitude in choosing model lists and great difficulty identifying directions.

Here, we extend the triangle example in several ways, varying the model list or the DGM to see how they interact under various deviations. Although we examine the variability given model lists, we do not investigate the uncertainty involved in of choosing a model list. In effect, this assumes that model list selection has already been done through a separate process. We suggest this can be done by invoking an information theoretic criterion as in Section 5.2. Consequently, we neglect the variability in the representativity of the model list for the model space. This limitation arises because our goal is to study the effect of model mis-specification on errors and the impact of model list mis-specification on various weighting schemes, not the effect of varying a model list on a specific weighting scheme.

On the other hand, we restrict our attention mostly to model lists that are of approximately the same complexity as the DGM. (The exceptions to this are where we wanted to evaluate the effect of complexity or needed the numerical properties of the functions to provide bounds.) Our notion of complexity is the same as the distance used in the triangle example, simple minded term-counting. This is unsatisfying because the DGM may differ from a model on the list by a small but very complicated function. We ignore this case because such careful identification of a DGM will often be impossible given the data at hand. Indeed, the most realistic cases will have DGM's that are a little more complicated than the elements of the model list and this degree of complication will be of size similar to other terms in the DGM and be representable by a relatively simple functional form.

An adaptive approach to model list selection would use the data to generate model lists, generate stacking and BMA averages, and sequentially discredit models that were too far wrong. However, without understanding the metric geometry of model lists and DGM's and the tradeoff between bias and variance in model lists it is premature to propose adaptive techniques for model list selection.

We investigate the interaction between DGM's, model lists, and averaging techniques in a series of computed examples. These illustrate principles we anticipate will be important for model list choice. This is done in the context of linear models with an independent normal error and roughly comparable variability in the main quantities. Section 2 is an extensive discussion of model averaging in general, including descriptions of the stacking techniques and BMA that we use. Section 3 presents four computed 'geometric' examples and Section 4 presents two 'typical' examples in an effort to map out a comparison between stacking and BMA. Section 5 provides some interpretation of our results in terms of optimality properties, information theory, and potential methodology. Finally, Section 6 reviews our overall conclusions.

2. Model Averaging – Bayes and Non-Bayes

A paradigmatic formulation of the problem of model averaging is given by Juditsky and Nemirovski (2000). Their setting is one which we have found convenient to adopt. Briefly, consider a compact, convex set $A \in \mathbb{R}^m$ with elements $\alpha = (\alpha_1, \ldots, \alpha_m)$. For now, assume A is contained in the L_1 ball, so that $\sum_{i=1}^m |\alpha_i| \leq 1$. Fix some function space and an element f in it. Choose a list of models, say $f_1, \ldots f_m$ in the space. Let f_A denote the best approximation to f using linear combinations of the f_i's weighted by the entries in the optimal element $\alpha^* \in A$. That is, we have defined $f_A = \sum_{i=1}^m \alpha_i^* f_i$

where $\alpha^* = (\alpha_1^*, ..., \alpha_m^*)$ is defined by

$$\alpha^* = \arg\min_{\alpha} \int \left(f(x) - \sum_{i=1}^{m} \alpha_i f_i(x) \right)^2 d\mu(x). \tag{2.1}$$

This optimality criterion is merely one of many which we could have chosen.

The overall goal of model averaging is the following. Given a set A, a list of functions $f_1, ..., f_m$, n observations of the form $Y_i = f(X_i) + \varepsilon_i$ for $i = 1, ...n$ assumed IID mean zero, common variance σ^2, and a bound L so that $\max(|f_i|, |f|) \leq L$ find an estimator as close to being as good as f_A as possible.

Two comments about this structure are appropriate here.

First, note that this assumes the f_i have been given so that implicitly the model list uncertainty is zero. However, we want to consider how well the model list approximates the true model. The key issue is how to choose the f_i's given that we do not know f. A secondary question is how to choose the model selection principle, or here, more specifically, the model averaging technique. After all, we have no guarantee that f can be represented as a sum of the f_i's. This leads to a notion of model list approximation error, separate from the model approximation error of how well a single model approximates the true model.

Second, the f_i's chosen are written as if they were functions in a function space. Often this is true. More generally, they can be regarded as parametrized functions, that is $f_i(x) = f_i(x|\theta)$. Examples of this include the case that the f_i's are neural networks with different architectures or linear regression models with different functions and possibly different explanatory variables. In these cases, there is an estimation problem for θ nested within each f_i as well as an estimation problem for α across the f_i's.

With these structures in mind we review several important contributions to the general approach of model averaging. We begin with the two techniques we use here: Stacking and BMA. Then we turn to functional aggregation, agnostic learning, greedy approximation and data fusion.

2.1 Stacking

The main idea in stacking is to combine $f_1, ..., f_m$ by a cross-validation technique. The idea is that the models are 'stacked' in layers f_i with weights α_i. In particular, Wolpert (1992) and Breiman (1996a) described stacking as follows (see also Smyth and Wolpert, 1998, for many, more recent, references). Define vectors

$$z_j = (z_{1,j}, ..., z_{m,j}) = (f_1^{-j}(x_j), ..., f_m^{-j}(x_j))$$

in which the superscript $-j$ means that the j^{th} observation (y_j, x_j) is not used to estimate the coefficients in each f_i which are then evaluated at the delete x_j. The α is chosen to minimize

$$L = \sum_j (y_j - \sum_i \alpha_i z_{i,j})^2. \tag{2.2}$$

There are various choices for the set A leading to different techniques for optimizing L and numerous variants on leave-one-out cross-validation. (Breiman, 1996a, argues that stacking works well in practice even though it has not yet been shown to satisfy an optimality principle formally.)

Our work differs from Breiman (1996a) because we are testing how well the method works in the presence of model approximation error. The procedure we followed was to consider a supermodel $Y = X\beta + \varepsilon$ and identify m submodels of it to play the role of the f_i's. Write these models as

$$Y_i = X_i\beta_i + \varepsilon_i,$$

for $i = 1, \ldots, m$. Fix $i = 1$ and use 4/5 of the data to estimate $\hat{\beta}$ and the remaining 1/5 of the entries in X_i with $\hat{\beta}_i$ to get a vector of fitted values $\hat{Y}_{(i=1),1/5}$, of length $n/5$. Doing this for each of the other fifths of the data in turn gives a vector of length n that we denote $Y_{(1)} = (\hat{Y}_{(1),1/5}, \ldots, \hat{Y}_{(1),5/5})$. Do the same for each of the other values of i to get $Y_{(1)}, \ldots, Y_{(m)}$, one vector of fitted values for each model. Note that delete-1 cross validation would use $(n-1)/n$ of the data in each (X_i, Y_i) for $i = 1, \ldots, m$ rather than $4/5$. We used fifths because it gives a better evaluation of the performance of the model in a predictive sense while retaining most of the data for estimating the parameters.

Now, we estimate β by the posterior mean. We have fixed $\sigma = 1$ and used a $N(0, \sigma)$ prior on each of the β's. It remains to determine the stacking coefficients by minimizing $||Y - \sum_i \alpha_i Y_{(i)}||_2$. There are several ways to do this. The first form of stacking is to permit α to be unconstrained, i.e., it ranges over $\mathbb{R}$. We call this $S1$. The second form of stacking we use is $S2$ in which we impose $\sum_i \alpha_i = 1$. In both cases negative coefficients are permitted. Since this is counterintuitive, we also consider a third and fourth form of stacking, denoted $S3$ and $S4$. In $S3$, we start with the $S2$ weights and impose $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$ by replacing negative α_i's with zero and then renormalizing so the nonzero α_i's sum to one. A more sophisticated approach replaces the truncation by a quadratic optimization that incorporates the constraint $\sum_i \alpha_i = 1$. This is done in $S4$. Indeed, it is seen that the techniques of stacking increase in their complexity and conceptually get closer to BMA, even though the determination of the $\alpha' s$ remains like cross-validation.

In any of these 4 stacking procedures our predictor is of the form

$$\hat{Y} = \sum_i \alpha_i (X_i \hat{\beta}_i) = \sum_i \tilde{X}_i (\alpha_i \tilde{\hat{\beta}}_i) \tag{2.3}$$

in which the tilde's indicate that the quantity under them has been lifted up to have dimension equal to that of the corresponding quantity in the supermodel by putting zeros in as necessary when an explanatory variable is not included in prediction. We do this so that the true β will have the same dimensionality as the overall estimate of β.

In evaluating the performance of the averaging technique we look at $R = R_n = E||\beta_T - \hat{\beta}||^2$ in which $\hat{\beta}$ is $\sum_i \alpha_i \tilde{\hat{\beta}}_i$ and $\hat{\beta}$ is an estimate based on n data points. We argue that this is the right analog to the Bayes approach which would be optimal and use posterior probabilities of models in place of the α_i's. Our use of R here is as a Bayes risk on the parameter estimates because we have chosen new parameters at random on each iteration. In fact, as observed in Section 5.1, if the entries in the matrix X are independent, this Bayes risk is a predictive criterion as well as an assessment of error.

2.2 Bayes Model Averaging

Again, we suppose $f_i(x) = f_i(x|\theta_i)$. Suppose also that we have a priors across families and priors within families. Now, for a set S, and dataset D the probability that S contains the data generating model, DGM, is

$$W(S|D) = \sum_{i=1} \int w(M_i, \theta_i|D) I_{f_{i,\theta_i} \in S} d\theta_i = \sum_{i=1} \int w(M_i|D) w(\theta_i|D, M_i) I_{f_{i,\theta_i} \in S} d\theta_i. \tag{2.4}$$

Within each integral the first posterior probability is the posterior for the i^{th} submodel and the second is the posterior density for the parameter in the submodel.

This leads to an average over models of the form

$$Y_B = \sum_i W(M_i|D) X_i E(\beta_i|D) \tag{2.5}$$

for the linear regression case, parallel to Equation 2.3. The general properties of BMA have been studied extensively. A few of the most important papers, among many others, include the following. Madigan and Raftery (1994) verified that BMA beats out model choice under a logarithmic scoring rule in that BMA provides better predictive ability than using any one model, perhaps because of its optimal treatment of what we call here the within model list uncertainty; see also the examples by Hoeting et al. (1999, Section 7). Clyde (1999) addresses some of the prior selection questions and model search strategies. She directly confronts the problem of the model space being too large to permit an exhaustive search. To deal with model uncertainty, she implements the orthodox Bayes program by a stochastic, rather than an algorithmic, search. George and Foster (2000) address the model uncertainty problem in Bayes model selection by an empirical Bayes technique which they related to conventional model selection principles. For further references, see Clyde (1999).

2.3 Other Non-Bayes Averaging Techniques.

Juditsky and Nemirovski (2000) developed an approach called functional aggregation. It uses the structure at the beginning of this section. They establish an upper bound on the error of approximating f by linear combinations of the f_i's. In their proof they estimate inner products between the f_i's which are much like covariances. This gives a method of construction resulting in an estimator satisfying their theorem. Stacking can also be expressed in terms of quantities that are like covariances (Smyth and Wolpert, 1998).

This setting has been elaborated on by Lee, Bartlet and Williamson (1996). Their technique, agnostic learning, satisfies an optimality criterion somewhat like Juditsky and Nemirovski (2000). However, their setting is much more general. They only assume a joint probability model for the X's and Y's and approximate the probabilistic relationship by a function within a general class of functions so as to minimize the expected value of a loss function. Their technique is primarily intended for neural networks. They still establish a theoretical bound on the performance of their method. In the proof of that theorem they introduce quantities that can be recognized as averages of models. Indeed, their function f^* is expressible in terms of partial sums denoted f_k. See also work by Haussler (1992) and Kearns and Vazarani (1994).

Another approach is through greedy approximation (Jones, 1992, 2000). The idea here is approximate a function in an L^2 space by another function within a subset of that space by evaluating it at a linear combination of the explanatory variables. Finding the best linear combination at one stage leaves a residual to which one applies the procedure again. At each stage one optimally fits the residual. The result is a sequence of partial sums of functions evaluated at linear combinations of explanatory variables. The partial sum converges to the true function. The main theorem establishes the rate of this converges in L^2 norm in terms of sample size.

Another recent approach is called boosting. The central idea here is that combining a large collection of weak prediction rules (through weighted majority voting for instance) is much easier that finding one highly accurate stand-alone prediction rule and still gives a very accurate prediction rule. A good overview, with a large reference list, can be found in work by Schapire (2002). In a sense, what is going on here is that the variability over prediction rules is substituting for data and modeling we wish we had.

A variant on the idea of boosting is called bootstrap aggregating or bagging. The earliest clear statement of the technique seems to be by Breiman (1996b). The idea is to enhance the performance of a predictor by repeatedly evaluating the predictor on bootstrap samples and then forming an average over those samples. Breiman (1996c) studied instability and in a separate paper (Breiman, 1996b) argued that bagging will produce substantial improvements in predictors when they are based on 'unstable' procedures such as neural nets, CART's, and subset selection in linear regression.

Despite the success of boosting and bagging – and the philosophically attractive centrality of prediction they have – we have focussed on stacking (a version of delete-1 cross validation) and BMA for two reasons. One is that these two methods have been well studied theoretically and computationally from the optimization standpoint so a comparison to seek where each works best seems timely. The second is we wanted to get at the DGM directly. Even though any good prediction scheme will be consistent for the DGM, we wanted to motivate the choice of model list and look at how well it can be used to approximate the DGM by different techniques. In principle, the decision rules one combines in a boosting context are analogs of the choice of model list but the linkage between boosting and function approximation is not as close. Also, it is well recognized that bagging (when it works) mostly reduces variance, not bias (Breiman, 1996b; Buhlman and Yu, 2002), and our interest here is on bias. (We comment that the collection of unfortunate directions identified in the introduction as having high rates of change may correspond, in a heuristic sense, to the unstable predictors where bagging seems to be most effective.)

A general approach by Clarke (2001) sought to combine model selection principles on sets of models on which they might be optimal. There it was proved that model averaging never does asymptotically worse than model choice for predictive purposes, if the model selection principles are consistent. This technique was implemented computationally by de Luna and Skouras (1999).

Finally, Luo and Tsitsiklis (1994) have an approach called data fusion. They combine functions from different sources in a communications network to get one overall message, or output. Their setting is information theoretic, but the procedures are similar.

3. Stacking vs. BMA: Geometric Examples

Here we present a collection of geometric examples to compare the performance of the four stacking procedures to BMA. We describe them as geometric because they stem from an effort to visualize the metric geometry of an infinite dimensional function space. Since these spaces are infinite dimensional, there will be diverse ways to draw two dimensional diagrams of multidimensional subspaces. Our diagrams are not unique; they are merely efforts to visualize the collection of functions that would be well approximated by a given model list.

For instance, Figure 1 depicts three functional forms as points. It would be equally valid to regard them as vectors coming out of the origin and forming some kind of pyramidal shape. The location of B and C relative to these points would be depicted differently in these two cases. Using linear independence of functions to form basis elements gives a nearly Euclidean geometry different from a supremum norm geometry which is better for some approximation purposes. Ultimately, the same collection of functions is represented but the properties of ther representation may be very different. Our term counting measure is cruder than these.

Alternatively, one can regard the models on the model lists that we use here as subsets of the elements of a basis for the function space containing the DGM. An individual term is a basis element. If the basis is ordered by some criterion like frequency, then we are evaluating how well averaging techniques perform using sets of basis elements that represent truncations of expansions in the basis.

Implicit in this is the view that selection of model lists will rely, at least initially, on choosing subsets of a basis based on geometric approximation principles. To this end, the names of the examples have been chosen to suggest the principle they illustrate.

3.1 Computational Results

We contrast stacking and BMA in four scenarios. The first three have a fixed DGM and several model lists. The fourth has fixed model lists and considers a class of DGM's. In all cases, the DGM is not on the model list and we evaluate the Bayes risk R by simulation. We report our approximate R's to two decimal place accuracy even though in some cases the precision is higher. We did this because an improvement of less than 0.01 is too small to prefer one method over another. Thus, even though a bigger number of replicates would give higher precision we only chose a number large enough, that a normal test for the differences between the two methods would reject equality (at the 0.05% level or better); in most cases we report 100. We always verified the results qualitatively for higher replication sizes. In addition, we chose sample sizes n for R_n large enough to permit reliable comparisons yet small enough to be reasonable in practice; in most cases we report 50. We always verified the results qualitatively for higher sample sizes.

We comment that we have presented a representative selection of all the cases we computed. In fact, in various places we give slightly stronger interpretations than are justified strictly on what we give here. In those places we indicate what we omitted.

3.1.1 DENSITY

Consider the case that the DGM is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

with ε's IID $N(0,\sigma^2)$. Suppose there are three model lists,

$$\mathcal{M}_1 = \{X_1^2, X_2, X_3, X_4, X_5; X_1, X_2^2, X_3, X_4, X_5; X_1, X_2, X_3^2, X_4, X_5\}$$

$$\mathcal{M}_2 = \mathcal{M}_1 \cup \{X_1, X_2, X_3, X_4^2, X_5\}$$

$$\mathcal{M}_3 = \mathcal{M}_2 \cup \{X_1, X_2, X_3, X_4, X_5^2\}.$$

Here, the terms separated by semi-colons are in the same model. For brevity we have not written in the β's which are random. Later, in Subsection 4 when we vary the DGM, we will drop the β's in the DGM too since the list of explanatory variables is all that is needed to specify the model. Clearly, none of the model lists contains the DGM $\{X_1, X_2, X_3, X_4, X_5\}$.

For 100 repetitions and a sample size of 50 we have the following table of Bayes risks. We have used an asterisk to indicate the entry in a column with the smallest risk.

It is seen that in all cases stacking wins over Bayes. As the model list index increases the risk decreases so it is the coarsest stacking procedure that wins.

Technique	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
BMA	0.37	0.29	0.24
S1	0.28*	0.24*	0.21*
S2	0.31	0.26	0.22
S3	0.31	0.25	0.21
S4	0.31	0.25	0.21

Table 3: Risks and model list density

As suggested by Figure 2, the models on the lists, denoted A, B, C, D, and E look like the positive octant of a sphere in five dimensions. Note that lines with the same number of marks are equal in length. That is, each of the models is two terms distance from the DGM at the center and any two models are four terms distant from each other. This means that if we consider a sphere $B(DGM, 2+\eta)$ centered at the DGM with radius $2+\eta$ for some $\eta > 0$ then $\text{card}\{\mathcal{M}_i \cap B(DGM, 2+\eta)\}$ is increasing in i, even when normalized by $\text{Vol}(B(DGM, 2+\eta))$ to give the approximate density of $\mathcal{M}_i$ at the DGM.

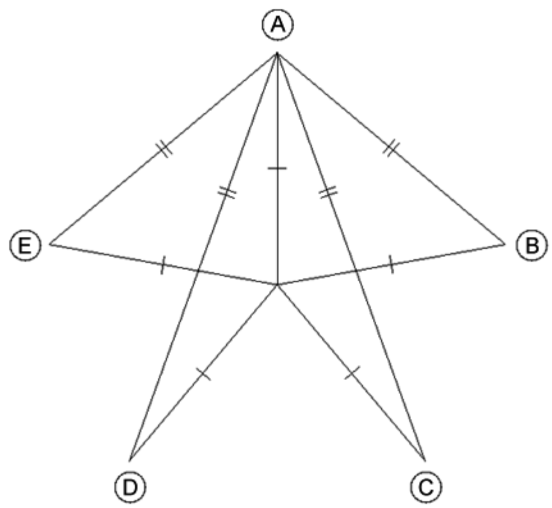

Figure 2: Density of the Model List Near the DGM.

We comment that in the simulation results presented here, we have assumed that all of the data comes from the same wrong model (although the β varies). In fact, one can redo the simulations so that, say, any proportion comes from the wrong model and the rest comes from one or more models in $\mathcal{M}_1$. As the proportion of data from a model on the list increases, the degree by which BMA outperforms stacking increases, in terms of R. We justify this formally in a short Appendix.

3.1.2 Bracketing

Here we argue that as the ability of the model list to provide bounds on the DGM increases, the degree by which stacking should outperform BMA also increases. This differs from the concept

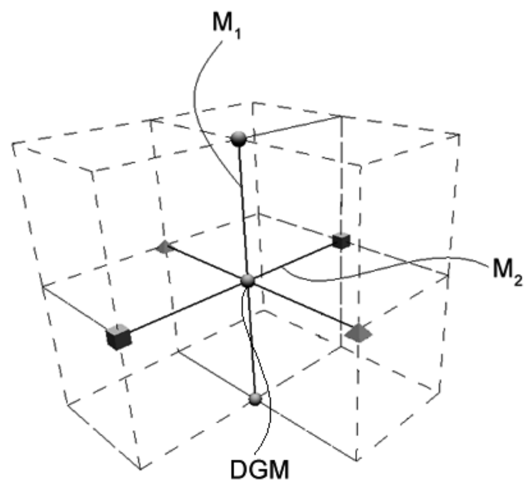

Figure 3: Interiorness.

of density (models per unit volume in the model space) because we are examining proximity in a function approximation sense. Given two models m_1 and m_2 which bracket a DGM, we imagine a region in model space that joins m_1 and m_2 and passes through the DGM. This is denoted by a line with endpoints indicated by spheres in Figure 3. Doing this for another pair of models that brackets the DGM gives another line ending in cubes in Figure 3. The plane formed from them represents the region in model space generated by the Cartesian product of the list consisting of the four models together. Thus, it is geometrically reasonable to represent the DGM as an interior point of any one of three regions in the model space: two lines and a plane. The Cartesian product of the plane with the line ending with pyramids would form a three dimensional interior if we continued to nest model lists. The intuition here is that as the DGM is situated in nested regions of increasing 'dimension', formed from larger and larger model lists whose elements bracket the DGM, the available volume near the DGM that the averaging strategy tries to fill up also increases. As this 'dimension' increases it is harder for two models to be close, a situation that should favor stacking over BMA.

Thus, we consider a simple model

$$Y = X_1 + X_2 + \varepsilon$$

where ε is normal as before and define

$$\mathcal{M}_1 = \{X_1 + \sqrt{|X_2|}\text{sign}(X_2); X_1 + X_2^2\},$$

$$\mathcal{M}_2 = \{\sqrt{|X_1|}\text{sign}(X_1) + X_2; X_1^2 + X_2\},$$

where sign$(\cdot)$ is the sign of its real valued argument and

$$\mathcal{M}_3 = \mathcal{M}_1 \cup \mathcal{M}_2.$$

Here, $\mathcal{M}_3$ is"two dimensional" in the sense that the first 2 model lists can be regarded as lines joining the two elements in $\mathcal{M}_1$ and $\mathcal{M}_2$.

Our comparison of the model spaces and model averaging techniques is summarized in the following table. We have used 100 simulations repetitions and a sample size of 50. (The pair of asterisks in the bottom row means that the two entries are equal, and smallest. In some simple cases, the optimizations that give S3 and S4 coincide.)

Model List	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
BMA	2.53	2.21	0.98
S1	2.55	2.22	1.35
S2	2.55	2.23	1.36
S3	2.46*	2.15*	0.84*
S4	2.46*	2.15*	0.88

Table 4: Risks and dimension of the interior

There are two implications. First, the risks decrease as the internal dimension of the manifold formed from the model list increases, when the extra dimensions added are functionally helpful. As indicated in Figure 3, if we define another model list with two models, we can imagine the DGM as an interior point of a three dimensional manifold. In results not shown here, if the extra dimension is generated by models which bound the DGM as a function of its explanatory variables then the risk decreases. However, adding more dimensions to $\mathcal{M}_3$ can increase the risk if the extra dimensions correspond to functions that cannot be used to bound the DGM or are otherwise strikingly different from it. Thus, the 'direction' here may be quantified as the difference between the DGM and approximations to it formed from models on the model list. Essentially, this is the bias, as in Section 1.

Here we argue that, subject to constraints, increasing complexity is helpful. Suppose the DGM is a sum of all second order terms in 3 variables,

$$Y = X_1^2 + X_2^2 + X_3^2 + X_1X_2 + X_2X_3 + X_1X_3 + \varepsilon.$$

Note there are six terms. We define three model lists each with 6 models in which each model differs from the DGM by four terms, the same sense of distance as used before. The model lists differ in their complexity, also as measured by number of terms: The models on list 1 have 8 terms, the models on list 2 have 6 and the models on list 3 have 4. Figure 4 shows one model R_i from each of the three model lists $\mathcal{M}_i$ in relation to the DGM. The lines joined to each R_i indicate the terms that get summed to form the model. Explicitly, the first, higher complexity, model list is

$$\mathcal{M}_1 = \{X_1^2, X_2^2, X_3^2, X_1X_2, X_2X_3, X_1, X_2, X_3; X_1^2, X_2^2, X_1X_2, X_2X_3, X_1X_3, X_1, X_2, X_3;$$

$$X_1^2, X_2^2, X_3^2, X_1X_2, X_1X_3, X_1, X_2, X_1X_2X_3; X_1^2, X_2^2, X_3^2, X_3X_1, X_2X_3, X_1, X_3, X_1X_2X_3;$$

$$X_1^2, X_3^2, X_1X_2, X_2X_3, X_1X_3, X_1, X_2, X_1X_2X_3; X_2^2, X_3^2, X_1X_2, X_2X_3, X_1X_3, X_1, X_3, X_1X_2X_3\}.$$

To form this list we dropped one term from the DGM but added three others in various ways. The list of equally complex models is

$$\mathcal{M}_2 = \{X_3^2, X_1X_2, X_2X_3, X_1X_3, X_1, X_2; X_1^2, X_1X_2, X_2X_3, X_1X_3, X_1, X_1X_2X_3;$$

$$X_1^2, X_2^2, X_3^2, X_1X_2, X_1, X_2; X_1^2, X_2^2, X_3^2, X_1X_3, X_1, X_1X_2X_3;$$

$$X_2^2, X_3^2, X_1X_2, X_2X_3, X_1, X_2; X_1^2, X_3^2, X_1X_3, X_1X_2, X_2, X_3\}$$

To form this list we dropped two terms from the DGM but added two in various ways. The list of lower complexity models is

$$\mathcal{M}_3 = \{X_1^2, X_2^2, X_3^2, X_1; X_1X_2, X_2X_3, X_1X_3, X_1; X_2^2, X_3^2, X_1X_2, X_1;$$

$$X_1^2, X_2^2, X_2X_3, X_1X_2X_3; X_3^2, X_1X_2, X_2X_3, X_1; X_2^2, X_1X_2, X_1X_3, X_2\}.$$

To form this list we dropped three terms from the DGM but added one in various ways. In all three cases, the new terms added were chosen to be relatively far from each other, in an effort to 'fill out' the model space.

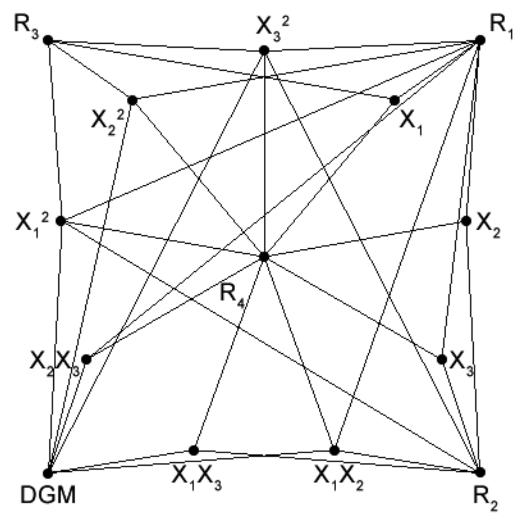

Figure 4: Effect of Complexity.

3.1.3 COMPLEXITY

We compared the performance of the four forms of stacking and BMA for these three model lists with 200 repetitions and a sample size of 100. These larger numbers arose because there were more terms in these models than in the earlier cases. However, as before, other numbers of repetitions and sample size were qualitatively the same. In the table below, when one of two equal entries is starred it means the two entries are equal to the exactitude shown, but later digits favor the starred one.

In each row, the risk increases from left to right as the number of terms decreases. Thus, for fixed distance, higher complexity helps. Moreover, the coarser stacking wins with the lower complexity, the more refined stacking wins with the higher complexity. This suggests that BMA would win with high enough model complexity (as measured here). This is consistent with the view that as more terms are included in the models, the model lists span a larger subspace thereby approximating a DGM better. Indeed, the amount by which BMA loses to the best of the stacking methods increases as the complexity decreases as seen in the last row. In view of the interpretation of the term-counting distance, one can regard this as a sort of 'maximum entropy' principle.

Model List	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
BMA	0.19	0.74	1.61
S1	0.15	0.29*	0.56*
S2	0.15	0.36	0.73
S3	0.14*	0.57	1.21
S4	0.14	0.44	1.03
BMA -Best	0.05	0.46	1.04

Table 5: Complexity

3.1.4 SENSITIVITY

Here we investigate the effect of varying the DGM while using one of two reasonable model lists. The DGM is varied by perturbing the exponents on explanatory variables. This is an important case because exponents are often chosen for convenience on the basis of scatter plots with unexamined variability. Also, there is a tradeoff: At what point does a deviation in the exponent change from a robustness check to a new explanatory variable?

Consider a collection of DGM's of the form

$$Y = X_1^{\alpha} + X_2^{\beta} + X_3^{\gamma} + \varepsilon,$$

with the usual linear model conventions. We choose a collection of values for the vector (α,β,γ) starting with $\alpha = \beta = \gamma = 1$ and then considering variants on it. The other triples were as follows. First we set $(\alpha,\beta,\gamma) = (4/5,4/5,4/5),(2/5,2/5,2/5)$ to represent the general effect of the explanatory variables entering by lower powers, then we set $(\alpha,\beta,\gamma) = (6/5,6/5,6/5),(8/5,8/5,8/5)$ to represent the effect of higher powers. Then we set $\alpha = 1$ so see how the other powers affected the results. We did computations for $(\beta,\gamma) = (4/5,6/5),(6/5,6/5),(4/5,4/5)$ and then for $(\beta,\gamma) = (2/5,8/5),(8/5,8/5),(2/5,2/5)$.

The two model spaces were

$$\mathcal{M}_1 = \{X_1 + X_2 + X_3; X_1^2 + X_2^2 + X_3^2; X_1X_2; X_2X_3; X_1X_3\}$$

and

$$\mathcal{M}_2 = \{X_1, X_2, X_3, X_1^2; X_1, X_2, X_3, X_2^2; X_1, X_2, X_3, X_3^2; X_1, X_2, X_3;$$
$$X_1^2, X_2^2, X_3^2, X_1X_2, X_2X_3, X_1X_3\}.$$

Thus, in terms of explanatory power $\mathcal{M}_1$ is equivalent to $\mathcal{M}_2$. However, the models in $\mathcal{M}_2$ are closer to each other in the term-counting distance than the models in $\mathcal{M}_1$ are, because the models in $\mathcal{M}_1$ have no overlapping terms.

To get a high enough precision we used 400 repetitions with sample size 100 to get the following table when the DGM was $\alpha = \beta = \gamma = 1$.

These results are expected: When the DGM is in on the model list BMA wins. (No patterns were noted for stacking risks.) However, in other cases, we found that even relatively small model approximation errors made BMA lose out to stacking. Here is one case with 100 reps and sample size 50. The other scenarios and sample sizes were qualitatively the same.

Sensitivity:	$\alpha=\beta=\gamma=1$		$\alpha=1,\beta=4/5,\gamma=6/5$	
Model List	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_2$
BMA	0.043*	0.043*	6.25	6.25
S1	0.068	0.059	6.24	6.35
S2	0.065	0.055	6.25	6.32
S3	0.412	0.043	4.53*	6.22
S4	0.045	0.048	6.03	6.17*

Table 6: Sensitivities for different values of α, β and γ

Thus, even with relatively small changes in the exponents stacking won and the finer methods of stacking tended to perform better. It is seen that the risks for the stacking procedures were all smaller for $\mathcal{M}_1$ than for $\mathcal{M}_2$. We suggest, like Breiman (1996a), this occurs because the elements of $\mathcal{M}_2$ have more terms in common than do the elements of $\mathcal{M}_1$ so they are closer together. Since having two exponents 0.2 away from 1 is enough to let stacking win over BMA, BMA is just not very robust. Indeed, in other series of computations involving changes in powers or truncated variables BMA performed very poorly relative to stacking.

4. Stacking vs. BMA: A Typical Setting

Next we examine all subsets regression to see how stacking and BMA compare in what many regard as a typical setting. In effect, we regard each term as a model rather than grouping terms into models. When we use all subsets of a collection of explanatory variables, interest focuses on determining which variables to include rather than finding appropriate functions of them. In some cases like this, BMA sometimes does better than stacking, possibly because the model list in this case is so much richer than in the geometric examples.

We treat two cases. The first, ‘Absolute’, case is in the same spirit as the earlier geometric examples. In the second, ‘Relative’ case, we reweight the perturbation term so it will have the same degree of variability and the other terms, regardless of functional form. This changes the results significantly although it will be difficult to do in practice because the perturbation terms are typically unknown. In the unnormalized case, like those before, the size of the model approximation error appears to dictate which of BMA and stacking will have lower risk. In the other, normalized, case it is the direction of model approximation error that matters.

4.1 Computational Results

Here we investigate the effect of using all subsets with 4 explanatory variables. Now, for $\{X_1;X_2;X_3;X_4\}$we use

$$\mathcal{M} = \{X_1;X_2;X_3;X_4;X_1,X_2;...;X_1,X_2,X_3,X_4\}.$$

as our model list and we consider two types of deviation term.

4.1.1 THE ABSOLUTE VERSION

We consider four possible DGM's:

$$Y_1 = X_1 + X_2 + X_3 + X_4; Y_3 = X_1 + X_2 + X_3 + X_4 + X_1^2;$$

$$Y_2 = X_1 + X_2 + X_3 + X_4 + X_1X_2; Y_4 = X_1 + X_2 + X_3 + X_4 + X_1^2 + X_1X_2.$$

Recall that X_1^2 is typically a bigger bias term than X_1X_2. The computational results are summarized here for 100 repetitions and a sample size 50.

	DGM 1	DGM 2	DGM 3	DGM 4
BMA	0.14*	1.20*	1.61	2.66
S1	0.27	1.39	1.82	3.04
S2	0.25	1.33	1.81	3.03
S3	0.78	1.81	2.08	3.05
S4	0.14	1.22	1.58*	2.66*

Table 7: All subsets

If we redo the computations dropping X_4 or $X_4 + X_3$ the results are qualitatively the same, although the risks in each entry decrease as the variables are dropped.

The pattern is clear: As the deviation term increases, the risks increase. In addition, BMA works better with small deviation terms like zero and a product; stacking works better with larger deviation terms, like higher powers. This relatively good performance by BMA is limited by the richness of the model list and the convenient form chosen for the DGM. Moreover, in columns 1 and 4 S4 and BMA are indistinguishable, given the tolerance. This held for other sample sizes and repetition numbers. Because the bias is relatively small, more sophisticated stacking techniques performed better than the rest.

4.1.2 A RELATIVE VERSION

In contrast with the absolute version we offer a relative version that gives contrasting results. The model list is the same as before but we now consider 4 different DGM's. They are

$$Y_1 = X_1 + X_2 + X_1^2; Y_2 = X_1 + \sqrt{|X_1|}\text{sign}(X_1);$$

$$Y_3 = X_1 + X_2 + X_1X_2; Y_4 = X_1 + X_2\chi_{X_2>0}.$$

These were chosen partially for variety, since the last example revealed the behavior of stacking and BMA for the present model list with deviant cross and square terms. We wanted some comparability and some novelty.

The key difference here is that we modify the prior on the coefficient of the perturbation term. This modification increases or decreases the variance of the perturbation term so that the overall variance of the term is the same as the other terms. For instance, for the first DGM in earlier examples we would have used $\beta_3X_2^2$, both β_3 and X_2 distributed independent $N(0,1)$. Here, we replace X_2^2 by $X_2^2/\sqrt{2}$. Now $\text{Var}(X^2/\sqrt{2}) = 1$, the same as the explanatory variables in the other terms. If we didn't do this, it would correspond to using a $N(0,2)$ prior on β_3: The expression

$\beta_3 X_2^2/\sqrt{2}$ with β_3 and X_2 distributed as independent $N(0,1)$'s corresponds to the expression $\alpha_3 X_2^2$ with α_3 distributed as $N(0,1)$. While this makes the variability equivalent, it reduces the effect of the functional form and presumes knowledge about the perturbation term we would never have. (If we did have such knowledge we would build it into the model list.)

For brevity, we only present and discuss some of results for S3, S4 and BMA since S1 and S2 never won. As before, we use $n = 50$ as our sample size. The results are, in an average risk sense, that BMA did best for cases 1 and 3. In these cases, S4 did almost as well as BMA, and S3 was distinguishably worse than S4. In cases 2 and 4, S3 did best. BMA did almost as well, and S4 was distinguishably worse. Thus, for nice, smooth deviations BMA wins, while for nonsmooth deviations the robustness of stacking gives better performance.

To probe this, we redid the simulations partitioning the outcomes into ten subsets based on the absolute value of the coefficient on the perturbation term. Then we calculated the corresponding average risks. The ratios of risks presented here by decile track how rapidly the performance of BMA falls off as the effect of the perturbation term increases. Note that as the decile increases the standardized DGM is further and further from the domain on which BMA on the model list would be optimal. Our results are in the following tables. A value greater than 1 means BMA is performing better than the stacking procedure indicated. A value less than 1 indicates BMA is performing worse.

Decile	1	2	3	4	5	6	7	8	9	10
S3/BMA	2.85	2.17	1.87	1.57	1.55	1.43	1.31	1.06	1.16	1.02
S4/BMA	1.12	1.07	1.06	1.07	1.04	1.03	1.05	1.01	1.03	0.99

Table 8: Ratios of risks for deviant square term

Decile	1	2	3	4	5	6	7	8	9	10
S3/BMA	2.44	1.45	1.37	1.07	1.03	0.93	0.86	0.79	0.75	0.71
S4/BMA	1.22	1.18	1.16	1.05	1.02	1.06	1.02	0.99	1.00	1.00

Table 9: Ratios of risks for deviant signed square root term

Decile	1	2	3	4	5	6	7	8	9	10
S3/BMA	2.55	1.97	1.88	1.79	1.47	1.43	1.28	1.13	1.15	1.04
S4/BMA	1.12	1.05	1.05	1.06	1.02	1.04	1.03	1.03	1.02	0.98

Table 10: Ratios of risks for cross term

It is seen that the standardization shrinks the perturbation effect of the square so that BMA does better, in contrast to the absolute case. Also, as before, BMA does better than stacking with the cross term. However, with the other two deviations, BMA only outperforms the stacking procedures until the size of the coefficients is in the 6th decile beyond which S3 performs best. It may be that beyond the 6th decile the perturbation is large enough that the higher sensitivity of BMA to model

Decile	1	2	3	4	5	6	7	8	9	10
S3/BMA	2.21	1.45	1.33	1.15	1.04	0.99	0.92	0.84	0.80	0.74
S4/BMA	1.22	1.18	1.16	1.04	1.01	1.06	1.02	1.00	0.99	0.99

Table 11: Ratios of risks for characteristic function term

mis-specification becomes a problem. That this occurs for some perturbation terms and not others shows that the optimality of BMA is dependent on the direction of the deviation.

It seems that S3 begins to win over BMA when the perturbation term has a coefficient greater than the median of the other coefficients and the direction of perturbation is 'outside' the model list in the sense that a square root or a truncation is not well modeled by a higher power. BMA tends to win, here, when the perturbation is in a direction readily accommodated by the model list – here, cross terms and higher powers – and is not too big. Taken together, it seems that when there is one perturbation term, direction matters more under normalization and size matters more without.

5. Relation of Results to Optimality Properties

In all but one of geometric cases where model approximation error was nonzero, the error was large enough that some form of stacking always performed better than BMA in terms of risk under squared error loss on the parameters. The only exceptions occurred in the bracketing setting when we used an extra model list having models that were functionally very different from the DGM. (We conjecture this occurred for the same reason as in scenario D in the triangle example.) On the other hand, BMA tended to outperform stacking for model lists large enough that the lesser robustness of BMA didn't matter. Here, we focus on the big, typical picture.

5.1 Bayes and Non-Bayes Optimality

Intuitively, one expects data driven methods like stacking to converge to their limits slower than methods such as BMA that are more dependent on likelihoods. In addition, one expects stacking to be more robust against model mis-specification. This property is important unless one model on the list is exactly right or so close to being exactly right that other sources of variability – such as model approximation error – are huge by contrast. However, when this is not the case the risk behavior of a non-Bayes method like stacking, i.e., leave-one-out cross-validation, will perform better.

Recall that the Bayesian concept of model uncertainty depends on the model list. The models on the list are assigned weights in $[0,1]$, even though one must assume one of the models is true to get optimality results. In these cases, all but one of the weights will go to zero and the last will go to one. On the other hand, when the DGM is not on the list the estimator formed by weighting elements from the list will typically converge to the member of the list closest to the DGM. Here, even though we rechoose the parameters within the model in each iteration, it is unclear that the posterior probabilities in the BMA will converge to numbers in $[0,1]$ to identify the mixture of models closest to the DGM. That is, instead of the BMA converging to an element of the closed convex hull of the model list, it might be only to an element of the model list. Since BMA loses its optimality as the DGM deviates from the model list, we want to know how big the deviation must be before another procedure, like stacking, works better.

To understand this it helps to think about the relevant optimality properties. BMA is the Bayes action under a squared error criterion. Also, the usual optimality of Bayes methods follows from the complete class theorem (Robert, 1997). However, this doesn't apply here because we have found risks with respect to a joint likelihood outside the model list which yielded the Bayes model average and this quantity is different from the Bayes risk used in the complete class theorem. In the absence of this optimality, it is no surprise that BMA doesn't perform better than stacking.

By contrast, stacking uses coefficients derived from a cross validation technique which is optimal under some predictive criteria. Indeed, delete-one cross-validation, as a model selection principle, is equivalent to AIC, Mallow's C_p, and a form of generalized cross validation (Li, 1987). Moreover, Shibata (1981) showed that AIC is asymptotically optimal for choosing the number of terms to include in a linear model when the dimension of the model is allowed to increase. (Hannan and Quinn 1979 establish this for a dependent case as well.) Most importantly here, Shao (1997) and Li (1987) show that AIC, and procedures equivalent to it such as delete-one cross-validation, is optimal in some predictive contexts. On the other hand, there are cases where delete-k cross-validation can be similar to the BIC, see Shao and Tu (1996) for more details.

Here, the risk we have used is equivalent to a predictive criterion in some cases. More formally, when the explanatory variables are independent and the DGM is on the model list both cross-validation and BMA are optimal because our Bayesian criterion has a predictive interpretation. The argument is as follows. Suppose there are p predictors in total. Fixed data yields $\hat{\beta}$, which is $p \times 1$, as the model-averaged estimate of β, which is also $p \times 1$. Instead of evaluating predictive performance at one or a few points in the design space, we want to average over the whole space. Let Z be $p \times 1$ and have the same distribution as gives rise to the X vectors. So, still for a single fixed data set, we can regard $E\big((Z^t * \hat{\beta} - Z^t * \beta)^2\big)$ as the prediction error averaged over the design space. But this can be rewritten as $(\hat{\beta} - \beta)^t * \mathrm{Cov}(Z) * (\hat{\beta} - \beta)$ which can then be averaged across $(\beta, data)$ to get a Bayes risk measure of performance. So far in our simulations, the covariate vectors are taken to be p independent standard normals, so that $\mathrm{Cov}(Z)$ is the $p \times p$ identity matrix, and we are simply averaging $(\hat{\beta} - \beta)^t * (\hat{\beta} - \beta) = ||\hat{\beta} - \beta||^2$ across $(\beta, data)$ realizations. (If we were to try dependent covariates we would have to include $\mathrm{Cov}(Z)$.)

This argument assumes independence of the explanatory variables. which does not hold in many of our examples. Nevertheless, in an important context we have a predictive interpretation.

Because cross-validation has an optimality property over a larger domain than BMA, we expect cross-validation will work better than BMA when its optimality supersedes Bayes optimality. This happens off the domain of Bayes optimality because one expects Bayes optimality on a small region is likely to be stronger (on that region) than the optimality of cross-validation which holds over a much larger region. Indeed, the two senses optimality, for cross-validation and for Bayes, are incompatible so one will not reduce to the other, in general. On the other hand, it is tempting to think of the model average from stacking as corresponding to a Bayes model average where each coefficient in the Bayes average has a separate prior rather than requiring the coefficients to form a probability on the model list.

This provides a partial explanation for the greater sensitivity of BMA to the model list than stacking has: When the hypotheses of Bayes optimality are not satisfied, for instance when the DGM is not an element of the model list, the drop off from the optimality of Bayes methods, as guaranteed by the complete class theorem, for instance, may be so rapid that BMA is regularly beaten by other procedures, such as stacking, that are less sensitive to the model list. This drop off may be important after relatively small deviations from the model list so that more data driven methods should be preferred. In fact, the more rapid convergence of BMA on the span of the model list may become harmful on its complement. Succinctly, BMA is not robust against deviations of the model.

However, recall that the key strength of BMA is the optimal way it makes use of the model list. That is, the sensitivity of BMA to the model list is not an argument against BMA so much as it is an argument in favor of a proper assessment of model uncertainty, model list uncertainty, and model approximation to determine when the sensitivity of Bayes to the model list is a problem, in part because of bias.

5.2 Methodological Implications of the Geometry of Model Lists

In the cases presented here, and in many others not presented, we found that even with relatively small model approximation error BMA loses out to stacking. Also, our results suggest coarser forms of stacking do better when the elements of the model list are more distinct or the deviation of the DGM from the model list is greater. It is seen that our results suggest choosing single term models and then averaging does better than using models with several terms. Breiman (1996a) suggested that the biggest gains arise when dissimilar sets of functions are used and our results are consistent with this. We suggest this may be due to an interaction between the coefficients in the averaging across models and the estimation of the coefficients within models.

Work by George (2001) to counteract dilution is also an effort to overcome excessive similarity among models. Dilution is the phenomenon that the posterior probability that should accumulate at a DGM can be spread over a set of wrong models that are close to the DGM. If this set is large and its elements are badly located relative to the DGM then the posterior probability at a model near the DGM can be smaller than the posterior probability at a model further from the DGM. Eliminating dilution is one reason why model averaging techniques in general tend to outperform model selection procedures, in particular for predictive purposes.

Taken together with our present work these considerations suggest the following. We want the model list to reflect the right domain in the model space. The domain should be as big as possible so we include all reasonable candidates yet should be as small as possible so we have manageable BSV. Moreover, we want to choose a model list to span the domain with elements as far apart as possible, subject to generating the same span. Clearly, in a predictive context such as we have used here, the span should decrease as data accumulate.

An adaptive procedure will satisfy these constraints. Start with a relatively large model list, whose members remain relatively distinguishable among themselves, spread out over the model space. To get predictions one should use a data driven method such as stacking. As data accumulate, isolate a region of the model space for elaboration. At stage 2, choose another model list, with members differing by one or two terms as most, within that region. Then, switch to BMA for this smaller region. Obviously, one could iterate the first stage to narrow the region represented by the model list further to ensure the bias would be small enough that BMA could be expected to work well. (In principle this could be done by choosing model lists with members robust to anticipated classes of perturbations, such as truncations, oscillations, and different exponents.) The point is to use BMA only at the last stage where model list elaboration is complete. Thus, one uses each technique where it performs best.

5.3 Data Compression and Mutual Information

Model selection is the ultimate data compression because the model is a summary not just for the data one has got but for all the data one might potentially get. So, to formalize the adaptive method just described, at least conceptually, recall the existing machinery of data compression (Cover and Thomas, 1991, Berger, 1971, Blahut 1987). Now, let us regard the potential models as messages to be sent. Since this is a nonparametric space, full specification of a DGM will typically require infinitely many bits to get perfect precision. Thus, we choose a finite collection of canonical representatives. This corresponds to a model list. Sending one of these representatives will require only finitely many bits – corresponding to a pre-selected degree of precision – and we will choose the representatives so that whatever the DGM is, it will never be unacceptably far from at least one of them.

5.3.1 MODEL LIST SELECTION AS A RATE DISTORTION PROBLEM

The big task is to choose the canonical representatives. Unfortunately, it is impossible to formulate a universal recipe that gives an optimal choice of canonical representatives to use as our model list. However, there is a criterion that will let us evaluate how close a proposed collection of representatives is to optimality. This criterion is the rate distortion function, RDF. Evaluating it for a proposed model list will tell us how well that model list satisfies a reasonable data compression criterion. Here, we do not propose achievement of the RDF lower bound as the solution to the model list selection problem. We only observe that the intuition behind the RDF, and its heuristic properties, may encapsulate some key features of the adaptive model averaging and evaluation of variability that we have considered here.

The general problem of data compression is to speed transmission of the important information by permitting a controlled loss of less important information. In the paradigm case we imagine an n-vector of data X^n IID according to a probability P and we seek representatives $\hat{X}^n(1),...,\hat{X}^n(M)$ for some M that we set equal to 2^{nR} for some $R > 0$ that we will call the rate. This is the rate used in Shannon's rate distortion theorem. A future outcome vector X^n will be approximated by the representative $\hat{X}^n(i)$ closest to it. To measure distance, or distortion, we use $d_n(X^n, \hat{X}^n(i)) = (1/n)\sum_{j=1}^n d(X_j, \hat{X}_j(i))$ for some univariate distance d. Now, for every X^n we define $\phi(X^n) = \arg\min_i d_n(X^n, \hat{X}^n(i))$ so that ϕ gives the canonical representative closest to the data string obtained.

The goal is to find a set of canonical representatives of the right size, with $Ed(X, \phi(X)) \leq \Delta$ where $\Delta > 0$ is a pre-assigned tolerance. In principle there are many sets of such representatives; we want one that achieves Shannon's RDF

$$R(\Delta) = \min_{p(x|\hat{x}): Ed(X,\phi(X)) \leq \Delta} I(X; \hat{X}),$$

at least asymptotically, where $I(\cdot;\cdot)$ is the Shannon mutual information between the two arguments. The function $R(\Delta)$ is the minimal number of bits needed, on average, to represent a source symbol with distortion bounded by Δ. The achievability direction of Shannon's rate distortion theorem says that for $R > R(\Delta)$ and any $\varepsilon > 0$ there is a code with rate R and distortion less than $\Delta + \varepsilon$. A converse holds too.

Regarding the DGM as a message within a function space, like an X^n, to be compressed, Shannon's rate distortion theorem implies there will be a model list that nearly achieves the RDF. Indeed, a comment by Blahut (1987, p. 209, par. 2) shows that one way to achieve the rate distortion function lower bound is to choose elements that are relatively far from each other in the chosen distortion measure; this is consistent with preferring the elements on a model list to be fairly dissimilar.

A limitation of this approach, shared by other approaches, is that we cannot uncover the DGM exactly (outside of repeated efforts and ever more data). At best, we can only identify a representative for it, as an approximation. The distortion is an approximation error and the probability of error in decoding is like the uncertainty of model selection. An important conceptual difference between data compression and model list selection is that data compression optimality properties integrate over the samples space and therefore implicitly assume a scheme will be used repeatedly while in statistics we choose the model once. Nevertheless, the statistician can regard the integration over the sample space as a pre-experimental design criterion for finding a model list from which to choose.

5.3.2 Distortion as Number of Terms in the Normal Case

It remains to choose the distance d that gives the constraint in the RDF minimization. In the spirit of information theory, consider using the conditional Shannon mutual information as d. In general, the relative entropy of Kullback-Leibler distance, between two probabilities P and Q is

$$D(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} \mu(dx)$$

where μ is a common dominating measure and p, q are Radon-Nikodym derivatives with respect to it. The Shannon mutual information, SMI, is the relative entropy between a joint distribution $P_{X\times Y}$ and the product of its marginal distributions, $P_X \times P_Y$, written

$$I(X;Y) = D(P_{X\times Y}||P_X \times P_Y).$$

The pointwise conditional SMI, pointwise CSMI, is the SMI between the conditional distributions with the values of the conditioning variables specified, for instance

$$I(X;Y|Z=z) = \int p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \mu(dx,dy).$$

The CSMI itself is the expectation of the pointwise CSMI: $I(X;Y|Z) = \int p(z) I(X;Y|Z=z)dz$. The relative entropy is not a metric but has distance like properties such as defining a convex neighborhood base. Locally, the relative entropy does behave like squared error, which is a metric.

Although it is defined quite generally, we use the SMI here for normal error linear regression. It is easy to verify that for $Y = X\beta + \varepsilon \tilde{N}(X\beta, \sigma^2)$ and $Y' = X\alpha + \varepsilon' \tilde{N}(X\alpha, \sigma'^2)$ we have

$$I(Y;Y'|\alpha,\beta,X,\sigma,\sigma') = \frac{1}{2}\left(\log(\frac{\sigma'}{\sigma})^2 - 1\right) + \frac{\sigma^2}{2\sigma'^2}\left(1 + X(\beta-\alpha)\right)^2.$$

To develop an interpretation for this as a distance in the present setting suppose the matrix X is a single row with three parts: $X = (X_1, \ldots, X_{k_1}, X_{k_1+1}, \ldots, X_{k_2}, \ldots, X_{k_2+1}, \ldots, X_{k_3})$. The first part of the string corresponds to the explanatory variables that are only in Y, i.e., the corresponding α_i's are zero. The middle string corresponds to the explanatory variables that are only in Y', i.e., the corresponding β_i's are zero. The third string has the explanatory variables common to both models. These are the only variables for which the corresponding β's and α's are not zero.

Assume, as we have in our computations, that all the α_i's and β_i's are $N(0,1)$ and that Y and Y' are the same physically so that the common X_i's will have the same coefficients, at least in some average sense. Then, if $\sigma = \sigma'$ we have

$$I(Y;Y'|A,B,X) = \frac{1}{2}E_X\left(\sum_{i=1}^{k_1} X_i^2 + \sum_{k_1+1}^{k_2} X_i^2\right).$$

If the X_i's are independent and have a common distribution then this counts the number of X_i's the two models do not have in common. It is the cardinality of the symmetric difference on model space defined in the introduction and used throughout. It is seen that under this distance it doesn't matter whether the approximation error is due to inclusion of an extra term in the DGM or in the model list. Both give the same discrepancy.

We comment that the expressions for the CSMI can be extended to include cases where the explanatory variables are correlated. This would give a notion of distance substantially different from merely counting terms. We have not derived these expressions even though our computed examples sometimes include terms that are not independent. In addition, it is unclear how use of the CSMI could extend to non-regression settings. For the present, we accept these limitations because our focus is on developing the intuition for independent variable regression settings.

6. Conclusions

Our general point is that BMA is more sensitive to model approximation error than stacking is when the variabilities of the random quantities are roughly comparable. We have sought to characterize this via distance and direction in the model space. This led to our assertion that BMA will be outperformed by stacking when the bias exceeds one term of size equal to the leading terms in the model or when the direction of deviation has a different functional form (with higher variability) that the model list cannot approximate well. Providing an information theoretic interpretation of model selection as a data compression problem led us to justify our use of the number of terms as an appropriate measure of distance and was consistent with Breiman's (1996a) intuition.

Our geometric examples in Section 3 confirm that several intuitively reasonable properties of model lists are desirable. One is high density: More models in the neighborhood of the DGM is better than few models in the neighborhood of the DGM. Another is high interiorness: More models that bracket the DGM, in the sense of making it an interior point of a higher dimensional shape, corresponds to better predictive performance. Higher complexity, in the sense of number of terms, is better than lower complexity. This is subject to fixing the cardinality of the model list and the distance from the model list elements to the DGM. Fourth, making the elements of the model list relatively distinguishable rather than permitting overlap improves approximation power. Our treatment in the more typical setting, especially the standardized case, shows the importance of direction and suggests the extra variation from the bias matters most.

We emphasize the narrowness of the examples we have computed and the consequent limitations on the generality of our conclusions. For instance, we can see the necessity of requiring the variabilities of X's, ε's and β's to be comparable by considering the following cases. Suppose we have a model list with three models: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, $Y = \beta_0 + \beta_2 X_2 + \varepsilon$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ where the X's, ε's, and $(\beta_0, \beta_1, \beta_2)$ are independent $N(0,1)$'s but that β_3 is $N(0,\tau^2)$. If $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ is the DGM, the intuition developed here predicts that stacking beats out BMA when τ^2 is small enough. If the distribution of β_3 is replaced with

$N(0,1)$ then BMA will win over stacking. On the other hand, in some cases (not given here) where BMA wins over stacking but the model list does not contain the true model and the variabilities of the β's and X's are roughly comparable, then, usually, as the variance of ε decreases, stacking does better and better ultimately beating out BMA. That is, a decrease in σ^2 tends to amplify the effect of model mis-specification.

There are other general points. If a DGM is not on the model list then we get convergence of a model average to the convex combination closest to the DGM. Thus, the SE for a parameter (or model) measures variability around the wrong model and assesses how close we are to identifying it rather than the true model. This means we must explicitly examine Draper's (1995) BSV or the bias to get an accurate expression for the variability.

Also, the geometry of how the model lists are situated within the model space, and where the DGM is relative to them matters greatly when comparing model averaging schemes. If the DGM is not on the model list we want to know how well we have approximated it. A model choice approach amounts to identifying the model that is closest to the true model. However, one expects to do better than this if some kind of averaging is used. Thus, we want a model list that will permit the DGM to be well represented by some averaging procedure. The rate distortion function, and vector quantization techniques in general, may be one way to do this.

A general approach may be adaptive. One wants to choose a readily distinguishable set of models from a non-parametric model space, choose a small set of these which are closest to the stacking average and then elaborate models in that region, iterating the procedure, and switching to BMA in place of stacking at the last stage because the robustness concerns will have been reduced. In essence, one wants the diameter of the model list to decrease at a rate reflecting the accumulation of data and the cardinality of the model list to increase until the approximation error is one term or less, and is in a convenient direction.

The design issue remaining is in the geometry. For a list of models, $\mathcal{M} = \{m_1, \ldots, m_k\}$ let the diameter be $\delta(\mathcal{M}) = max_{i,j} d(m_i, m_j)$ for some distance d and real number δ. Now, we want a procedure for choosing a model list $\mathcal{M}_n$ at stage n so that $\#(\mathcal{M}_n)$ increases slowly and $\delta(\mathcal{M}_n)$ decreases slowly. Ensuring the sequence $< \mathcal{M}_n >$ is optimal and choosing the appropriate model averaging procedure at each stage may give better results in an adaptive context than any fixed method.

Acknowledgments

The author gratefully acknowledges the insightful discussions and programming genius of Paul Gustafson. Thanks also to Andrew Barron, Merlise Clyde, and Nando de Freitas for helpful discussions. Finally, the author thanks two anonymous referees whose detailed comments greatly improved this paper. This research was supported by NSERC Operating Grant 5-81506.

Appendix A.

Here we give the argument that using contaminated distributions induces a linear term in the error where the coefficient is related to the fraction of contamination.

Recall, we are repeatedly sampling from $(\mathcal{X}, \mathcal{M}, \Theta, Y)$, equipped with a distribution we denote generically by $f(x)f(m)f(\theta|m)f(y|\theta,m,x)$. Here, x is a realized value of the design matrix treated as a random variable $\mathcal{X}$, m is a realized value of $\mathcal{M}$, a random variable varying over the model list, θ is a realized value of Θ which we take to represent the parameter, and y is the outcome of Y, the response of interest. In our examples, $f(x)$ factors into a product of independent $N(0,1)$'s; the number of factors equals the number of entries in the overall design matrix. The model variable $\mathcal{M}$ varies over the model list. Here, we use one of two choices for the distribution of $\mathcal{M}$. The typical case is that $\mathcal{M}$ is degenerate: It assigns probability one to a fixed model, where model means a selection of explanatory variables (the entries in $\mathcal{X}$) and choice of corresponding factors, the β_i's. When this fixed model differs from the fixed model generating the data, the Y_i's, we say we are in the wrong model case.

More generally, we needn't choose a degenerate $\mathcal{M}$. Indeed, we can choose any distribution for the elements of the model list. The weight put on a specific model is the probability that we will obtain a data set assuming it to be true. The parameter vector Θ is the 'β-vector' in the linear regression model, possibly with the σ in the error term ε. Usually, the β_i's are independent $N(0,1)$ random variables. Finally, given the model, the parameter, and the design matrix, we generate Y as the response by adding a $N(0,\sigma^2)$ error term, ε.

The error we estimate by simulations is

$$E\left(SQE(\mathcal{X},\mathcal{M},\Theta,Y)\right) = E\left(\int \left[\hat{E}(Y|X=x) - E(Y|X=x)dx\right]\right)$$

where the first expectation in the integral (with the 'hat') depends on $\mathcal{X},Y$ and assumes $\mathcal{M}\tilde{f}(m)$, while the second expectation in the integral depends on $\mathcal{M},\Theta$. If $f(\mathcal{M})$ is a product of IID $N(0,1)$'s then $SQE = ||\hat{\beta}-\beta||^2$, where the β's have dimension equal to that of the full model.

When the data come from the right model, i.e., one in $\mathcal{M}_1$, then we evaluate

$$E_f\left(SQE(\mathcal{X},\mathcal{M},\Theta,Y)\right)$$

where f is the four-fold density above. In the wrong model case, we evaluate

$$E_g\left(SQE(\mathcal{X},\mathcal{M},\Theta,Y)\right)$$

where $g(x,m,\theta,y) = f(x)g(m)f(\theta|m)f(y|\theta,m,x)$ and g is degenerate at the wrong model (the other densities are the same as in f).

If we write

$$g_\eta(x,m,\theta,y) =$$

$$f(x)\left((1-\eta)f(m)+\eta g(m)\right)f(\theta|m)f(y|\theta,m,x) = (1-\eta)f(x,m,\theta,y)+\eta g(x,m,\theta,y)$$

then we see that

$$E_{g_\eta}\left(SQE(\mathcal{X},\mathcal{M},\Theta,Y)\right)$$

$$= (1-\eta)E_f\left(SQE(\mathcal{X},\mathcal{M},\Theta,Y)\right)+\eta E_g\left(SQE(\mathcal{X},\mathcal{M},\Theta,Y)\right)$$

so we see that as the proportion of data from the wrong model increases, the risk increases linearly in that proportion. Typically, BMA will win when the true model is in the model list. However, when the data comes largely or entirely from the wrong model our computations suggest stacking will typically give smaller risks. The degree of improvement depends on the geometry of the model list and where the DGM sits relative to that list.

References

Toby Berger. *Rate Distortion Theory.* Prentice-Hall, Englewood Cliffs, NJ, 1971.

Robert Berk. Limiting Behavior of Posterior Distributions When the Model is Incorrect. *Annals of Mathematical Statistics*, 37(1):51-58, 1966.

Richard Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, Reading, MA, 1987.

Leo Breiman. Stacked Regressions. *Machine Learning*, 24:49-64, 1996a.

Leo Breiman. Bagging Predictors. *Machine Learning*, 24:123-140, 1996b.

Leo Breiman. Heuristics of Instability and Stabilization in Model Selection. *Annals of Statistics*, 24(6):2350-2383, 1996c.

Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199-231, 2001.

Peter Buhlman and Bin Yu. (2002) Analyzing Bagging. *Annals of Statistics*, 30(4):927-961.

Merlise Clyde. Bayesian Model Averaging and Model Search Strategies. In *Bayesian Statistics 6*, pages 157-185, Valencia, Spain, 1999.

Bertrand Clarke. Combining Model Selection Procedures for Online prediction. *Sankhya A* 63:229-249, 2001.

Thomas Cover and Joy Thomas. *Elements of Information Theory.* Wiley, NY, 1991.

Xavier de Luna and Kostas Skouras. Model Metaselection. Technical Report 203, Dept. of Stat. Sci. UCL, 1999.

David Draper. Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society*, 57(1):45-97, 1995.

Edward George. Dilution Priors for Model Uncertainty. MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, California. Available electronically via `http://www.msri.org/publications/ln/msri/2001/nle/george/1/banner/01.html` 2001.

Edward George and David Foster. Calibration and Empirical Bayes Variable Selection. *Biometrika*, 87(4):731-747, 2000.

Paul Gustafson and Bertrand Clarke. Decomposing Posterior Variance. *Journal of Statistical Planning and Inference*, forthcoming.

David Haussler. (1992). Decision-Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Information and Computing*, 100(1):78-150, 1992.

Edward Hannan and Barry Quinn. The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society B*, 41(2):190-195, 1979.

Jennifer Hoeting, David Madigan, Adrian Raftery, and Christopher Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 4(4):382-417, 1999.

Lee Jones. A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *Annals of Statistics*, 20(1):608-613, 1992.

Lee Jones. Local Greedy Approximation for Nonlinear Regression and Neural Network Training. *Annals of Statistics*, 28(5):1379-1389, 2000.

Anatoli Juditsky and Arkadii Nemirovski. Functional Aggregation for Nonparametric Regression. *Annals of Statistics*, 28(3):681-712, 2000.

Zvi Kam. Generalized Analysis of Experimental Data for Interrelated Biological Measurements. *Bulletin of Mathematical Biology*, 64:133-145, 2002.

Michael Kearns and Umesh Vazarani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, MA, 1994.

Wee Sun Lee, Peter Bartlett, and Robert Williamson. Efficient Agnostic Learning of Neural Networks with Bounded Fan-in. *IEEE Transactions on Information Theory*, 42(6):2118-2132, 1996.

Ker Chau Li. Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15(3):958-975, 1987.

Zhao-Qing Luo and John Tsitsiklis. (1994) Data Fusion with Minimal Communication. *IEEE Trans. Inform. Theory* 40(5):1551-1563.

David Madigan and Adrian Raftery. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, 89:1535-1546, 1984.

Thomas Minka. Bayesian Model Averaging Is Not Model Combination. Available electronically at `http://www.stat.cmu.edu/ minka/papers/bma.html`, 2000

Christian Robert. *The Bayesian Choice: A Decision-Theoretic Motivation.* Springer-Verlag, New York, 1997.

Robert Schapire. (2002) The Boosting Approach to Machine Learning: An Overview. *MSRI Workshop on Nonlinear Estimation and Classification*.

Jun Shao. An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, 7:221-261, 1997.

Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap.* Springer-Verlag, New York, 1996.

Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45-54, 1981.

Padhraic Smyth and David Wolpert. An Evaluation of Linearly Combining Density estimators via Stacking. Technical Report 98-25, Information and Computer Science Department, University of Irvine, 1998.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. Technical Report, Dept. of Statistics, University of Toronto, 1994.

David Wolpert. Stacked Generalization. *Neural Networks*, 5:241-259, 1992.

Yuhong Yang. Combining Forecasting Procedures: Some Theoretical Results. *Econometric Theory*, forthcoming.

Yuhong Yang. Aggregating Regression Procedures for Better Performance. *Bernoulli*, forthcoming.

Yuhong Yang. Regression with Multiple Candidate Models: Selecting or Mixing? *Statistica Sinica*, forthcoming.