

An Empirical Approach to the Comparison and Combination of Model Selection Criteria

Andrew K. Smith

DRAFT, February 19, 2008

1 Introduction

Model selection remains, despite the considerable progress that has been made in this area, a fundamental and very challenging problem in virtually every area of applied statistics. A multitude of model selection criteria have been proposed in the literature to address this problem. Broadly, one can think of these criteria as being of three different types: some measure in-sample fit only, possibly with a penalty imposed on the number of parameters in the model. Many popular criteria, such as AIC, BIC, and Mallows' C_p fall into this class. Others use cross-validation to estimate out-of-sample prediction performance. The most common example from this class is PRESS in linear regression. Finally, others use a true holdout sample, not used in the construction of the model itself, as a measure of out-of-sample prediction error. One example of this type is MAD, or median absolute deviation, sometimes used in time-series analysis.

While all of the above types of criteria can certainly be useful, a natural question to ask is whether one can formulate a criterion which simultaneously considers more than one of these types of goodness of fit. No commonly-used criterion, for example, considers both in-sample fit and cross-validation error. In this chapter we propose a new procedure which generates a new class of combined model selection criteria. This procedure allows the analyst to combine, for example, the benefits of good in-sample fit as measured by criteria such as AIC or BIC, and also good out-of-sample prediction performance into a single criterion.

This new procedure generates a very large (in fact, infinite) new class of criteria, and therefore the question of how to compare different criteria is very important. In the model selection literature, criteria are typically compared on the basis of the proportion of simulated data sets in which the criterion chooses the known correct model from a set of candidates. We propose a generalization of this procedure based on ranks of criterion values which, we argue, is a more realistic measure of a criterion's usefulness in an applied context. The traditional method of comparing criteria turns out to be a special case of our more general comparison methodology.

Combining these two contributions, then, our main result is an algorithm which, as we show, can be proven to find the optimal combination of a fixed set of criteria, either using the traditional definition of optimality as above or a more general definition which we discuss below. Since the straightforward use of a single criterion is a special case of our combined criteria, our algorithm is a true generalization of the traditional model selection procedure in the sense that the optimal combined criterion can be no worse than any of the original criteria.

The rest of this chapter is organized as follows: In section 2 we propose our method to combine existing selection criteria via a simple ranking procedure. Section 3 then discusses a generalization of the traditional method of comparing criteria, and presents an algorithm to select an optimal combined criterion. In section 4 we present simulation results from our algorithm, focusing on the two special cases of ARIMA and linear regression models. We discuss the theory behind our algorithm, and in particular prove that the algorithm can find the optimal combined criterion, in section 5. Computational details are presented in section 6. Section 7 presents

a discussion of inferential issues involved in our algorithm, and proves the ϵ -optimality of the solutions produced. In section 8 we discuss the role that prior distributions play in our algorithm. Finally, in section 9 we conclude and present possible topics for future research in this area.

2 Combining Model Selection Criteria

We will assume throughout that we are working with a fixed set of candidate models

$$\mathcal{M} = \{M_1, M_2, \dots, M_s\},$$

where s denotes the cardinality of \mathcal{M} .

Let X denote the matrix of covariates, which we assume to be fixed, and let the response be denoted by \mathcal{M} . Associated with each model $M_i \in \mathcal{M}$, we assume that there is an equation

$$\mathbf{y} = g_i(\mathbf{X}, \boldsymbol{\theta}(M_i)) + \epsilon$$

The notation $\boldsymbol{\theta}(M_i)$ is used to indicate that the parameter spaces associated with each model may be different.

If we fix a particular model selection criterion, say BIC for the sake of example, then the model selection problem becomes an optimization problem of the form

$$\min_{\{i=1,2,\dots,s\}} BIC(M_i).$$

If we view the optimization from this perspective, combining several MSCs is problematic due to their different scales. For example, considering the sum

$$BIC(M_i) + MAD(M_i)$$

is meaningless because BIC is computed based on log-likelihoods, while MAD is on the same absolute scale as the original observations. Thus, even if one allows a linear combination of BIC and MAD, choosing appropriate constant multipliers to make the combination meaningful is a difficult task. The chief problem with this simple approach, then, is *scaling*.

However, a simple modification to the above procedure can make the linear combination meaningful. Rather than viewing the original problem as minimizing the absolute BIC value, we can view it as minimizing the *rank* of the BIC value of M_i among the set of BIC values of all the models in \mathcal{M} . Considering the rank of a model's MSC value rather than the absolute MSC value itself has the advantage of automatically putting all MSCs on the same scale. If we have several MSCs, say $MSC_1, MSC_2, \dots, MSC_k$, we can then form a meaningful combination of them

$$MSC_{\boldsymbol{\alpha}}(M_i) = \alpha_1 \cdot R_{MSC_1}(M_i) + \alpha_2 \cdot R_{MSC_2}(M_i) \dots + \alpha_k \cdot R_{MSC_k}(M_i)$$

for any vector of convex coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)^T$ and where $R_{MSC_j}(M_i)$ denotes the rank of $MSC_j(M_i)$ among the set $\{MSC_j(M_1), MSC_j(M_2), \dots, MSC_j(M_s)\}$. Note that the values of $MSC_{(1,0,0)}$ are not the same as those of the original criterion MSC_1 due to the rank transform, though the ordering of the values is the same. Notice that it is not necessary to consider linear combinations with positive weights other than convex combinations, since any linear combination with positive weights amounts to a re-scaling of a convex combination — that is, it is simply a convex combination multiplied by some constant. It is easy to see that multiplying the objective function by a constant does not change the optimal solution, and therefore the scaling of the set of linear coefficients is irrelevant, which justifies our restriction to considering only convex combinations. Linear combinations involving negative weights are possible in principle, but are quite unintuitive since we expect all model selection criteria to be

	AIC	BIC	PRESS	sum
Model 1	1	3	4	8
(*) Model 2	2	2	2	6
Model 3	3	4	1	8
Model 4	4	1	3	8

Table 1: Example of the potential utility of combined criteria: No individual criterion selects the correct model (Model 2), but the sum of ranks does.

of some value in distinguishing the true model from some of the false ones. This intuition is confirmed by our simulations below. Notice that this way of resolving the scaling problem noted above crucially depends on the specification of the set of candidate models, \mathcal{M} – in effect we are using the criterion values of all the models in \mathcal{M} as a way of imposing a scale. As a preview of the potential effectiveness of this method, we present a simple artificial example in Table 1 below. We present many more simulation results in Section 4.

However, the above method of considering ranks is not the only way to make different criteria have similar scale. Another simple alternative is to *standardize* the values of each criterion. This yields another class of combined criteria, which are very much related to, but not identical to, those produced by ranking. As we show below, however, these criteria tend to be less effective than those produced by ranking. Why this is the case remains an open question. We expect it is due to the distributions of the criterion values themselves – standardizing the values forces a common mean and variance, but does *not* guarantee any particular type of distribution, while ranking always produces values with exactly the same distribution. In the sequel we will focus primarily on the approach employing ranks, but we give some simulation results for the standardizing method in 4.

Armed with these new methods of constructing combined MSCs, we are now faced with the problem of finding the optimal convex combination of existing MSCs. Of course, in order to do so, one must specify what exactly “optimal” means with respect to an MSC. We address this question in the next section, and then in section 4 we propose an algorithm for computing the optimal criterion.

3 Comparing MSCs

In the model selection literature, criteria are often compared in a similar way – see, e.g., [3], [1]. Typically, for a fixed candidate model set \mathcal{M} , some data sets are simulated and all models in \mathcal{M} are fitted to the resulting data. Different criteria are then compared on the basis of how often each criterion chooses the correct model – that is, how often the model with the optimal criterion value is in fact the true model chosen in the simulation. In this section, we propose a much more general framework in which to compare criteria, which we argue can be more useful in applied situations.

In order to determine what makes one criterion better than another, one should consider the applied context in which criteria are used. An idealistic approach to model selection may consist of a procedure such as the following: Choose a model selection criterion (say, BIC), fit all the feasible models in \mathcal{M} to the data, and calculate the BIC for each model. Then select the model with the minimum BIC. The underlying assumption, of course, is that if the model selection criterion is well-designed, then the true model ought to be the one selected by the criterion. If one adopts this approach, then the existing method of comparison described above seems quite natural.

In practice, however, this approach is rather naive, as it amounts to the analyst allowing the criterion to completely *dictate* model choice, rather than to guide it. An experienced analyst would always use other considerations such as diagnostic plots along with criterion values in

order to construct a sensible model. A more realistic procedure would be to consider several models with small BIC, and then use other non-quantitative information, such as residual plots, to choose among these top few candidates. If one uses this procedure, however, the comparison method above becomes less meaningful – we are not particularly interested in the probability that our criterion chooses the true model, but rather the probability that the criterion includes the true model in the top few choices. It is this general model selection strategy which motivates our method to compare criteria.

The key new idea to our approach is to consider ranks. To illustrate, we need a bit more notation. Suppose we have fixed k “basis” MSCs and the candidate model set \mathcal{M} . With a vector α of convex coefficients (i.e., $\alpha \geq 0, \sum_{i=0}^k \alpha_i = 1$), let us define

$$R_{\alpha}(M_i) = \text{rank}_{MSC_{\alpha}}(M_i),$$

where the rank is taken with respect to the set $\{MSC_{\alpha}(M_1), MSC_{\alpha}(M_2), \dots, MSC_{\alpha}(M_s)\}$. With this notation, the traditional approach to comparing criteria described above corresponds to calculating $P\{R_{\alpha}(M^*) = 1\}$, where M^* is the true model, for different values of α . The best criterion would then be the one with highest such value.

However, defining the ranks as above allow us to compare criteria in a much more general way. Rather than considering only the binary outcome $R_{\alpha}(M^*) = 1$, we can regard $R_{\alpha}(M^*)$ as a random variable and consider any functional of its distribution. Indeed, there are many other functionals besides $P\{R_{\alpha}(M^*) = 1\}$ which may be meaningful. As suggested above, one useful alternative would be $P\{R_{\alpha}(M^*) > c\}$, where c is some specified constant. It is natural to think of c as a maximum number of models the analyst is willing to consider “by hand” – that is, the number of candidates suggested by the criterion from which the analyst is willing to choose based on information other than the criterion itself. Of course, c might vary depending on the type of problem. For example, model selection in time series is notoriously difficult using diagnostic plots – often one has considerable trouble distinguishing between a simple AR(1) and an MA(1) process based on plots of the autocovariance function. For other types of models such as linear regression, there are more diagnostics at our disposal, and c may correspondingly be larger in the hopes that our chance of finding the true model will accordingly be better. Other functionals such as the mean and median can also be used.

In general, we can define an arbitrary functional T of the empirical distribution of R_{α} , and formulate our optimization problem as

$$\min_{\alpha} T(\hat{F}(R_{\alpha}(M^*))).$$

We have implicitly assumed that $R_{\alpha}(M^*)$ may be treated as a random variable. This is most naturally interpreted in a Bayesian context in which we assume a full probability model for the data. Such a framework would consist of

1. $\pi_{\mathcal{M}}$, a prior distribution on the set of candidate models
2. $\pi_{\Theta_{\mathcal{M}}}$, a prior on the parameter space associated with each model in \mathcal{M}
3. $f(\epsilon)$, an assumed distribution on the errors of the model

With all of these ingredients, we can now formulate our main algorithm, listed as Algorithm 1.

4 Results

It is a well-known fact that BIC is the only *consistent* model selection criterion – that is, as the sample size grows, BIC is guaranteed to choose the correct model. Thus, by considering

Model	BIC	Cp	Adj.Rsq
y ~ 1	264.049	3.072	0.000
y ~ X1	268.527	4.995	0.010
y ~ X2	268.413	4.880	0.009
y ~ X3	265.864	2.341	-0.018
(*) y ~ X4	265.399	1.886	-0.023
y ~ X1 + X2	272.637	6.548	0.016
y ~ X1 + X3	267.835	1.838	-0.034
y ~ X1 + X4	269.247	3.199	-0.019
y ~ X2 + X3	268.816	2.781	-0.024
y ~ X2 + X4	269.694	3.633	-0.015
y ~ X3 + X4	269.891	3.825	-0.013
y ~ X1 + X2 + X3	271.722	3.203	-0.030
y ~ X1 + X2 + X4	273.800	5.198	-0.009
y ~ X1 + X3 + X4	272.358	3.808	-0.024
y ~ X2 + X3 + X4	273.334	4.746	-0.014
y ~ X1 + X2 + X3 + X4	276.062	5.000	-0.022

BIC	Cp	Adj.Rsq		$\alpha = (0.78, 0, 0.22)$		Ranked combined values
1	5	13		3.64		3
6	13	15		7.98		8
5	12	14		6.98		6
3	3	8		4.10		4
(*) 2	2	5		2.66		1
13	16	16		13.66		14
4	1	1		3.34		2
8	6	7	→	7.78	→	7
7	4	3		6.12		5
9	8	9		9.00		9
10	10	11		10.22		11
11	7	2		9.02		10
15	15	12		14.34		16
12	9	4		10.24		12
14	11	10		13.12		13
16	14	6		13.80		15

Figure 1: An illustration of Algorithm 1. The top panel shows the raw values for each of the 16 models in \mathcal{M} . The bottom panel illustrates the sequence of transformations – rank by column, combine the columns using convex coefficients, re-rank the resulting values.

Data: $\pi_{\mathcal{M}}$, $\pi_{\Theta_{\mathcal{M}}}$, $f(\epsilon)$, \mathcal{M} , \mathbf{X} , sample size N
Result: A vector of ranks of the true model
for $i = 1 : N$ **do**
 Choose j , the index of the true model, from $\pi_{\mathcal{M}}$;
 Choose $\theta(M_j)$ from $\pi_{\Theta(M_j)}$;
 Simulate errors from $f(\epsilon)$;
 Set $\mathbf{y} = g_j(\mathbf{X}, \theta(M_j)) + \epsilon$;
 Fit all models in \mathcal{M} to \mathbf{y} ;
 Compute the matrix \mathbf{B} of ranked basis criterion values;
 For each α , compute the rank of the j th element of the vector $\mathbf{B}\alpha$.
end

Algorithm 1: Optimal combination of model selection criteria

	X1	X2	X3	X4
X1	1.000	0.504	0.774	-0.545
X2	0.504	1.000	0.421	-0.036
X3	0.774	0.421	1.000	-0.854
X4	-0.545	-0.036	-0.854	1.000

Table 2: Correlation matrix of predictor variables used to generate Figure.

combinations of BIC with other criteria, we are in effect asking to what extent this asymptotic result holds in particular finite samples. In this section we explore this question in two applied contexts – regression models and ARIMA models.

4.1 Regression Models

Variable selection in linear regression is perhaps the oldest and most-studied model selection problem in statistics. Here we give a few examples of our algorithm applied to linear regression problems.

For the interesting case, the covariate matrix was generated as 4 independent standard normals, each of length 40, which was then right-multiplied by another random matrix to induce correlation among the predictors. The correlation matrix is given in Table 2. \mathcal{M} was simply the set of all 2^4 subsets of predictors from the matrix. The prior $\pi_{\mathcal{M}}$ was specified indirectly by giving randomly assigned weights to each model size, and the weight for each model size was split equally among all models of that size. $\pi_{\Theta(\mathcal{M})}$ and $f(\epsilon)$ were both standard normal distributions. The grid was generated using a mesh of 0.02, and the basis MSCs were BIC, Mallows Cp, and adjusted R^2 . We did 2 runs of 2000 simulated data sets each: on the first run, we used combined ranked criterion values, and in the second we used combined standardized criterion values. The results are displayed in Figure 2. Note that the improvement observed by allowing combinations of these MSCs is actually unexpected, since all 3 criteria are based only on penalized in-sample fit. Nevertheless, we find the results quite interesting due to the substantial improvement of the combined criteria. Also interesting is the high sensitivity of the response surface near the optimum. Indeed, the optimum is quite close to the BIC corner, but the objective value varies greatly in this small region.

It is also instructive to compare the results of the ranked values versus the standardized values. In general, the two response surfaces look very similar, as one might expect, due to the very high correlation between the raw values and the ranked values of random vectors in general. Furthermore, the optimal points are very close to each other on the two surfaces. The primary difference is the moderately worse performance of the standardized values when considering

the functional $P(R_{\alpha}(M^*) > 1)$, noting that the optimal value using standardized values is nearly 76%, while for ranked values it is under 74%. We saw this phenomenon in several similar experiments (results not shown here,) and expect that it is true in general. For this reason, we focus only on ranked values in the sequel.

4.2 ARIMA Models

Some example simulation results are given in Figures 3 and 4. Figure 3 is fairly typical – in particular, the optimal point for each summary function is either at or near one of the corners, indicating that combined criteria are of little use in this case. Further, the different summary functions all behave similarly. Figure 4, on the other hand, has a more interesting structure. Especially noteworthy is the fact that the optimum point for the function $P\{R_{\alpha}(M^*) > 1\}$ is not near any of the 3 corners, indicating that the convex combinations of criteria do better than any individual criterion in this case. Further, the improvement is quite substantial – over 5%. This example also illustrates the importance of considering which functional to consider, since the optimal criterion varies considerably depending on which function is chosen.

Interestingly, the only difference in the parameters used to generate Figures 3 and 4 is the specification of $\pi_{\mathcal{M}}$. Figure 3 used a uniform prior over all models, and Figure 4 used a prior giving weight only to models of size 3. Both figures were based on 500 simulated data sets of size 100, and \mathcal{M} was defined for both to be all ARIMA models up to order $(2, 1, 2)$. In both cases, all parameters were generated as independent $U[-1, 1]$ random variables, subject to the restriction that the resulting model be stationary, and the errors were independent standard normals.

We caution against drawing general conclusions based on the example results shown here. The behavior of the response function depends on a variety of other factors, such as the structure of the matrix of regressors, and the distribution of the error terms, in an extremely complex way which has not been fully explored. The results shown here are intended only to serve as examples, and should not be used to infer the relative merits of particular criteria or combinations of criteria for any specific problem. One should also note any conclusion on the overall utility of combining model selection criteria from the few examples given here. Indeed, we feel that one of the biggest advantages of our procedure is that it is easy to apply it to any problem, and in particular we need not rely on general recommendations. Rather, by the same simulation technique used to generate the example figures, we can obtain results tailored to the specific problem under consideration.

5 Theoretical Considerations

In this section we explore the theory behind our algorithm. In particular, we investigate the behavior of the functionals such as $\mathbb{E}[R_{MSC_i}(M^*)]$, which are the primary functions of interest. Our results show that these functionals are piecewise constant, discontinuous functions of α , and thus in a sense they justify our consideration of only a discrete set of values of α in Algorithm 1. To simplify the discussion, we will operate under the assumption that there are never any ties in the set of values of an individual MSC, e.g., there are never ties in the set $\{BIC(M_1), BIC(M_2), \dots, BIC(M_s)\}$. This assumption is generally justified because most MSCs involve log-likelihoods, and are thus continuous. Further, define \mathbf{B} as a random matrix formed by taking the vectors $[R_{MSC_j}(M_1), R_{MSC_j}(M_2), \dots, R_{MSC_j}(M_s)]^T, j = 1, \dots, k$ as its columns. Thus, by the above assumption, each column of \mathbf{B} is a permutation of the integers 1 through s . We have the following theorem which illustrates the role of the convex coefficients α in the distribution of \mathbf{B} .

Theorem 5.1 *Let α be a vector with all nonnegative entries, such that $\sum_{i=1}^k \alpha_i = 1$. Suppose that, for all possible values of the matrix-valued random variable \mathbf{B} , the values in the vector $\mathbf{B}\alpha$*

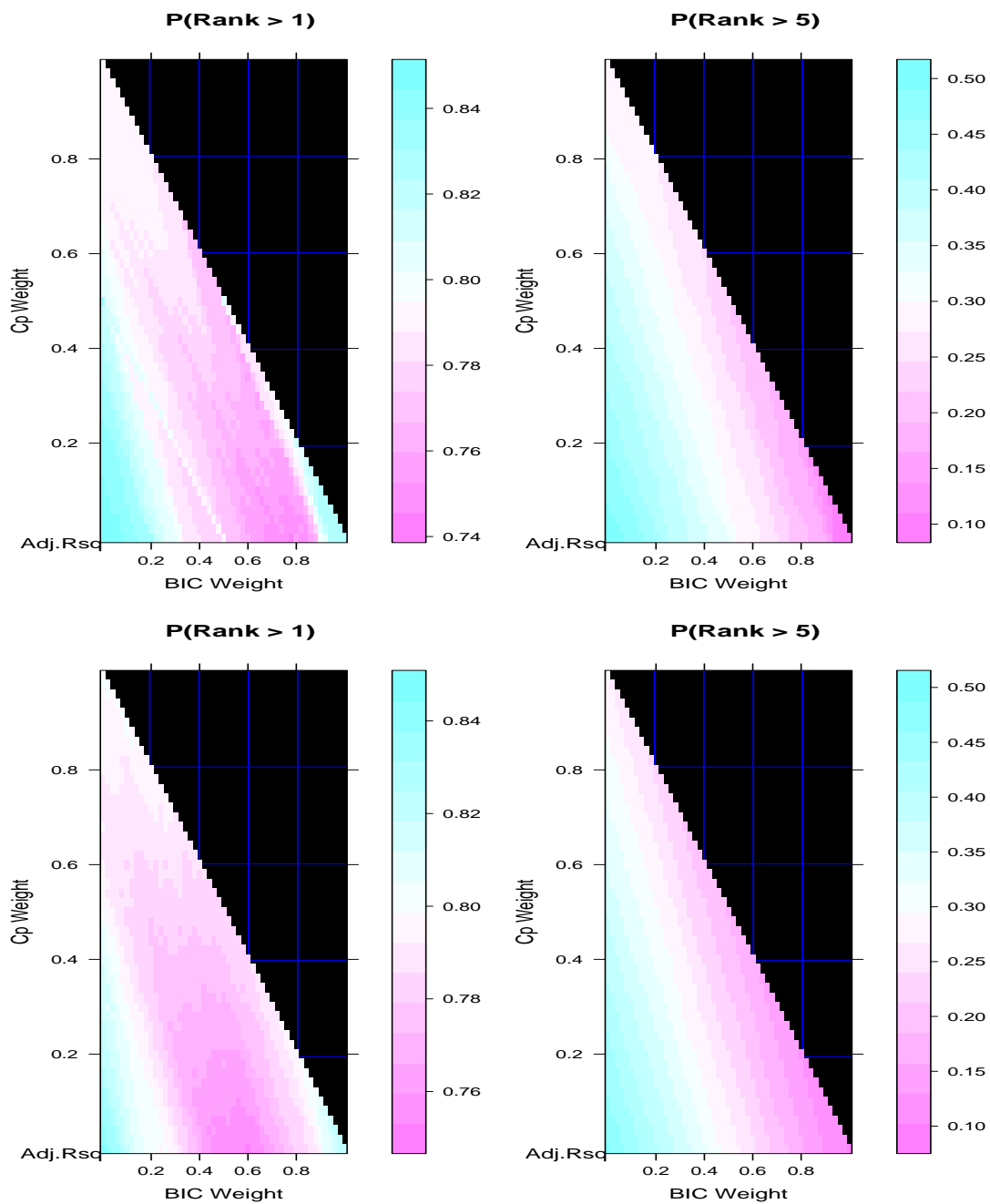


Figure 2: Example of regression in which combined criteria perform better than basis criteria. Top panels: Use of combined ranked basis criterion values. Bottom panels: Use of combined standardized basis criterion values. Note the slightly inferior performance of the standardized criteria. Figures are best viewed on a color display.

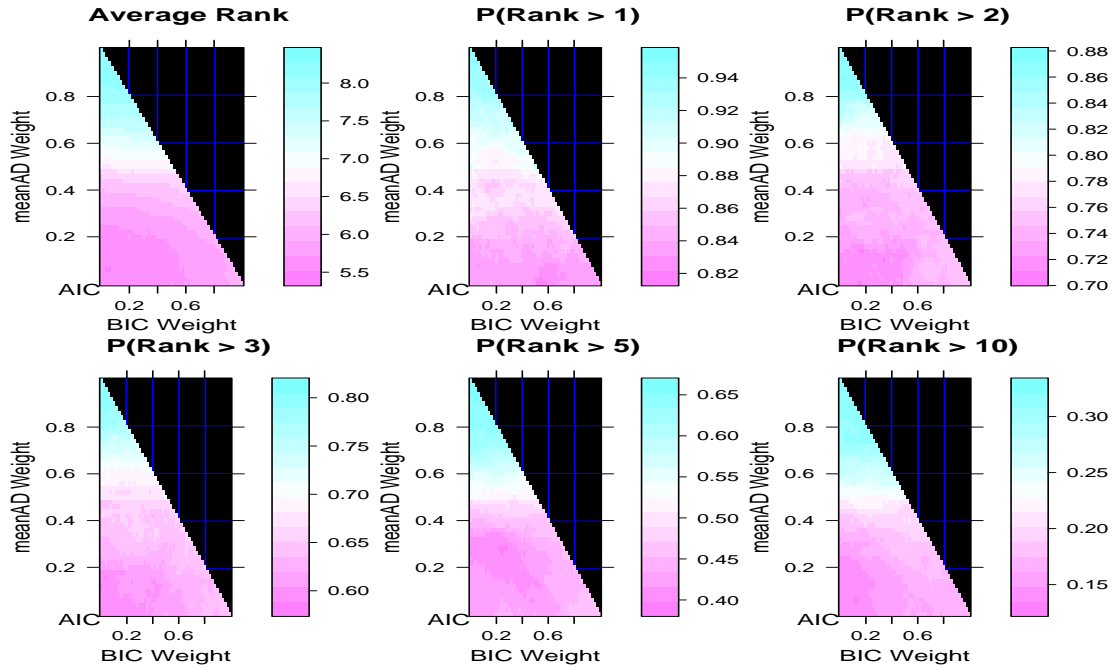


Figure 3: A fairly typical result for ARIMA models, in which the combined criteria appear to be of little value.

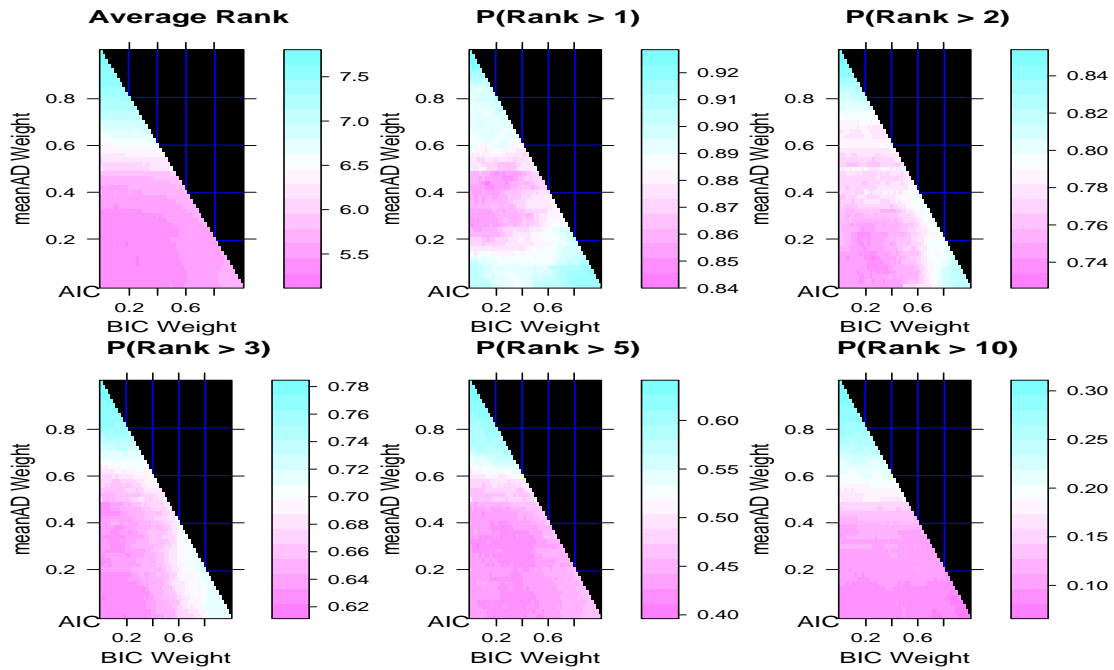


Figure 4: A more interesting result for ARIMA models, illustrating the potential utility of combined criteria.

are all distinct. Then there exists $\epsilon > 0$ such that

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| < \epsilon \Rightarrow \text{rank}(\mathbf{B}\boldsymbol{\alpha}) = \text{rank}(\mathbf{B}\boldsymbol{\alpha}')$$

Proof. First, note that by elementary linear algebra,

$$\|\mathbf{B}\|_2 \leq \|\mathbf{B}\|_F = \left(\sum_{i,j} \mathbf{B}_{ij}^2 \right)^{\frac{1}{2}} = \left(k \cdot \sum_{i=1}^s i^2 \right)^{\frac{1}{2}} = \sqrt{\frac{k \cdot s \cdot (s+1) \cdot (2s+1)}{6}}$$

Now, we have

$$\begin{aligned} \|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_\infty &\leq \|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_2 \\ &\leq \|\mathbf{B}\|_2 \cdot \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2 \\ &\leq \sqrt{\frac{k \cdot s \cdot (s+1) \cdot (2s+1)}{6}} \cdot \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2 \end{aligned}$$

Thus, by choosing $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2$ sufficiently small, we can make $\|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_\infty$ as small as desired. Now, let $c = \min_{\mathbf{B}, i \neq j} |(\mathbf{B}\boldsymbol{\alpha})_i - (\mathbf{B}\boldsymbol{\alpha})_j|$, where the subscript is used to denote the index of a component of a vector. Note that $c > 0$ since by assumption $(\mathbf{B}\boldsymbol{\alpha})_i \neq (\mathbf{B}\boldsymbol{\alpha})_j$ for any \mathbf{B} when $i \neq j$, and there are only a finite number of possible values of \mathbf{B} since each column must be a permutation of $[1, \dots, s]^T$. Suppose that we have chosen $\boldsymbol{\alpha}'$ such that

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2 \leq \frac{c \cdot \sqrt{6}}{2\sqrt{k \cdot s \cdot (s+1) \cdot (2s+1)}}.$$

Then, for each i , $|(\mathbf{B}\boldsymbol{\alpha})_i - (\mathbf{B}\boldsymbol{\alpha}')_i| \leq \|\mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\boldsymbol{\alpha}'\|_\infty \leq \frac{c}{2}$, and it is easy to see that $\text{rank}(\mathbf{B}\boldsymbol{\alpha}) = \text{rank}(\mathbf{B}\boldsymbol{\alpha}')$. \square

Theorem 5.1 naturally leads to the question of which values of $\boldsymbol{\alpha}$ satisfy the condition that there are no ties in $\mathbf{B}\boldsymbol{\alpha}$ for any \mathbf{B} .

Theorem 5.2 *Let k denote the number of base MSCs. Then the set $A = \{\boldsymbol{\alpha} : \mathbf{B}\boldsymbol{\alpha} \text{ has at least 1 tie for some } \mathbf{B}\}$ has $(k-1)$ -dimensional Lebesgue measure 0.*

Proof: For there to be a tie in $\mathbf{B}\boldsymbol{\alpha}$, the equation

$$B_{i1}\alpha_1 + B_{i2}\alpha_2 + \dots + B_{ik}\alpha_k = B_{j1}\alpha_1 + B_{j2}\alpha_2 + \dots + B_{jk}\alpha_k \quad (5.1)$$

must hold for some integers $i \neq j$. Now, since each column of \mathbf{B} has distinct entries, we know that $B_{mi} \neq B_{ni}$ when $m \neq n$. Thus we can rewrite (5.1) as

$$n_1 \cdot \alpha_1 + n_2 \cdot \alpha_2 + \dots + n_{k-1} \alpha_{k-1} + n_k \cdot \alpha_k = 0$$

where the n_i are all nonzero integers. Since the vector $\boldsymbol{\alpha}$ is assumed to be a vector of convex coefficients, we can use the condition $\sum_{i=1}^k \alpha_i = 1$, to write

$$(n_1 - n_k)\alpha_1 + (n_2 - n_k)\alpha_2 + \dots + (n_{k-1} - n_k)\alpha_{k-1} = -n_k$$

This is simply a linear equation in the $k-1$ variables $\alpha_1, \dots, \alpha_{k-1}$, and it is well-known that the set of solutions has dimension at most $k-2$, and therefore has $(k-1)$ -dimensional Lebesgue measure 0. Finally, note that since by assumption, each column of \mathbf{B} is a permutation of the integers 1 through s , there are only a finite number of possible values for \mathbf{B} . The result follows. \square

Since the set of convex combinations of k criteria is a $k - 1$ dimensional set, Theorem 5.2 tells us that for almost all convex combinations, there can be no ties in $\mathbf{B}\alpha$, and thus Theorem 5.1 applies. Thus, these two theorems together give us a qualitative description of the behavior of the response as a function of the convex coefficients — it is a locally flat step function. This runs contrary to the intuition one might develop from looking at the simulation results presented in Section 4.

If one is willing to disregard the “small” set of convex coefficients which may result in ties in $\mathbf{B}\alpha$, then Theorems 5.2 and 5.1 together imply that we need only check a finite number of values for α in order to find the optimum for any particular functional — it is easy to see that if we test a grid of small enough mesh, at least one point in the grid must lie in the optimum region.

6 Computational Considerations

The computation involved in our algorithm can be quite demanding, both in terms of storage and computing requirements. The computation required depends primarily on the size of \mathcal{M} , and the number of grid points chosen (which itself is a function of the grid size and the number of MSCs being combined.) There is a natural tradeoff when choosing the size of the grid — a smaller grid size provides better insight into the behavior of the cost functional as a function of the MSC weights, yet it can result in an exponential increase in both computation and storage requirements.

The relative amount of computation required for the fitting of all models in \mathcal{M} and for the calculation of all points on the grid depends on the type of model being fitted. For example, regression models are comparatively very easy to fit, and therefore in our simulations with regression models, most of the computation time is spent on the calculations for the grid. On the other hand, fitting ARIMA models requires a costly iterative maximization procedure, so the model fitting part of the algorithm typically dominates in ARIMA simulations. Deciding on an appropriate grid size and number of iterations, then, very much depends on the specifics of the problem.

Storage can also become an important issue, particularly when the number of MSCs to be compared is large and the corresponding grid of convex combinations becomes large in size. One key special feature to exploit is the structure of the summary functions we wish to compute for each convex combination. Essentially, the result of Algorithm 1 will be a vector of ranks for each combined MSC:

$$R_{\alpha}(M_i^*), i = 1, 2, \dots, N$$

If the summary functions can be computed dynamically, then we need not actually store this vector. Rather, we can update the summary function value and discard the individual ranks. Thankfully, this special structure occurs in many simple summary functions one might wish to use, including $\mathbb{E}[R_{\alpha}(M^*)]$, $\text{Var}[R_{\alpha}(M^*)]$, $P\{R_{\alpha}(M^*) > k\}$ for fixed k .

In choosing the number of base MSCs to use, there is of course a tradeoff between computation and storage requirements and the possibility of discovering more powerful combinations. Our experience indicates that using more than 3 or occasionally 4 MSCs is usually unhelpful. Further, it is also worth considering the similarities between MSCs. For example, many criteria are of the form

$$\text{loglik} - \lambda p$$

where p is the number of parameters in a fitted models. There is often little to be gained by considering many criteria of this same form. It is more likely that combinations of different *types* of criteria will be useful — e.g., an in-sample criterion of the form described above combined with an out-of-sample criterion based on a holdout sample. In our examples, we have always chosen 3 as the number of base MSCs, both to keep computational and storage requirements reasonable and also to allow easier graphical interpretation of the results.

7 Inference

In light of the simulation results presented above, there are several natural questions to ask:

- Is the observed improvement in performance by the combined criteria *significantly* better than the performance of the corresponding original criteria? That is, could the observed improvement be attributed to sampling error?
- Can we guarantee that the chosen combination is (nearly) optimal?

7.1 Hypothesis Testing

In the above simulation results, the empirical minimum of the response surface lies close to, but not exactly on, one of the corners of the domain, which naturally leads to the question of hypothesis testing: Is the observed minimum of the surface significantly smaller than a pre-specified value, or smaller than the values observed at the corners (which correspond to traditional model selection criteria)? This type of hypothesis allows one to determine if it is statistically worthwhile to consider the possibility of using our combined MSCs, rather than simple MSCs.

There are several possible approaches to this problem, each of which has some benefits and some drawbacks. The first is to regard the problem as a simple hypothesis test. Associated with each MSC we have a vector

$$(R_{\alpha}(M_1^*), R_{\alpha}(M_2^*), \dots, R_{\alpha}(M_N^*))^T,$$

where the subscript denotes the iteration number. This vector is a sample from the population of ranks of true models, under the priors $\pi_{\mathcal{M}}$ and $\pi_{\Theta(\mathcal{M})}$, and assuming the set of feasible models is \mathcal{M} . We are then free to apply any of a number of nonparametric tests directly to two pre-specified MSCs, say MSC_i and MSC_j . For example, if the cost function is $\mathbb{E}[R_{\alpha}(M^*)]$, we can directly compare the means of the two vectors by a simple paired t -test. Recall that we have required that the set of feasible models \mathcal{M} be finite, which guarantees that the observations $R_{\alpha}(M_l^*)$ are bounded for all α, l , which ensures that the populations involved have finite variances, and hence the CLT is always applicable, provided the sample sizes are sufficiently large. For comparing proportions of observations which exceed a fixed threshold k , a simple 2-sample proportion test can be used. For comparing the medians of two samples, there are simple nonparametric tests available. In particular, if we wish to test the hypothesis that the median of the sample differences is 0, we can use a sign test.

The difficulty with this approach is that it is invalid in the context of our original question, which is to test the observed *minimum* of the response surface against the original MSCs from which the combined MSCs were constructed. This is a case of the well-known problem of data snooping — in paying attention only to the minimum of the surface, we are implicitly testing many hypotheses simultaneously, which inflates the Type I error rate. The usual remedy to this problem of using the Bonferroni method or studentized range distribution to adjust α is worthless here due to the overwhelmingly large number of hypotheses being tested, which is usually in the hundreds or thousands. An easy though statistically inefficient solution is to split the sample into two parts. Using only the first part, we locate the minimum of the response surface, corresponding to, say, MSC_{α} for some particular value of α . We then compare the distribution of ranks corresponding to MSC_{α} to the control MSCs, typically the original MSCs such as AIC and BIC, using only the observations from the second part of the sample. Since the two parts of the sample are independent and we have pre-specified the hypotheses to be tested using the second part of the sample, we are *not* implicitly testing many hypotheses in using this procedure, so a significant result may be interpreted with more confidence. Though the power of such hypothesis tests are decreased due to the smaller sample size which we may use without “cheating,” we always have the option of simulating more observations. The only limitation is the computation involved in generating such observations.

7.2 Sample Size Estimation

To simplify notation, in this section we will abbreviate $R_{\alpha}(M^*)$ as R_{α} . If the cost functional has the special form

$$T(F(R_{\alpha})) = \mathbb{E}[G(\mathbf{B}, \alpha)]$$

for some function G , then the Algorithm 1 is actually a special case of the Sample Average Approximation method described in [2]. For the remainder of this section, we restrict our attention to this special case, noting that the functionals $\mathbb{E}[R_{\alpha}]$ and $P\{R_{\alpha} > k\} = \mathbb{E}[I\{R_{\alpha} > k\}]$ are both expectations of functions of \mathbf{B} and α .

Adopting the notation of [2], let \mathcal{S} denote the set of convex coefficients being tested. Further, let

$$\begin{aligned} v^* &= \min_{\alpha} \mathbb{E}[G(\mathbf{B}, \alpha)] \\ \mathcal{S}^* &= \{\alpha : \mathbb{E}[G(\mathbf{B}, \alpha)] = v^*\} \\ \mathcal{S}^{\epsilon} &= \{\alpha : \mathbb{E}[G(\mathbf{B}, \alpha)] \leq v^* + \epsilon\} \end{aligned}$$

and define sample counterparts \hat{v}^* , $\hat{\mathcal{S}}^*$, and $\hat{\mathcal{S}}^{\epsilon}$, where the expectation is replaced by sample average in the corresponding definitions. If we specify a type I error rate p and a positive number $\delta \in [0, \epsilon)$, then by Equation (2.23) in [2], we have that $\hat{\mathcal{S}}^{\delta} \subset \mathcal{S}^{\epsilon}$ with probability at least $1 - p$ if

$$N \geq \frac{\sigma_{\max}^2}{(\epsilon - \delta)^2} \cdot \log\left(\frac{|\mathcal{S}|}{p}\right), \quad (7.2)$$

where $\sigma_{\max}^2 = \max_{\alpha \in \mathcal{S} \setminus \mathcal{S}^{\epsilon}, \alpha' \in \mathcal{S}^*} \text{Var}[G(\mathbf{B}, \alpha) - G(\mathbf{B}, \alpha')]$. The most important feature of Equation 7.2 is that it depends only logarithmically on the size of the set of coefficients being considered, mitigating the exponential increase in the number of convex coefficients accompanying the addition of a basis MSC or a decrease in the mesh of the grid. We now derive a simple bound on the size of σ_{\max}^2 .

Theorem 7.1 1. If $G(\mathbf{B}, \alpha) = R_{\alpha}$, then $\sigma_{\max}^2 \leq (s - 1)^2$

2. If $G(\mathbf{B}, \alpha) = I\{R_{\alpha} > k\}$ for some constant k , then $\sigma_{\max}^2 \leq 1$.

Proof. For any α , the random variable R_{α} can take only the values $\{1, 2, \dots, s\}$. Therefore, for any α, α' ,

$$R_{\alpha} - R_{\alpha'} \in \{1 - s, 2 - s, \dots, -1, 0, 1, \dots, s - 2, s - 1\}.$$

Now, we have

$$\begin{aligned} \text{Var}[R_{\alpha} - R_{\alpha'}] &= \mathbb{E}[(R_{\alpha} - R_{\alpha'})^2] - \mathbb{E}[R_{\alpha} - R_{\alpha'}]^2 \\ &\leq \mathbb{E}[(R_{\alpha} - R_{\alpha'})^2] \\ &\leq (s - 1)^2 \end{aligned}$$

where the last inequality follows since $|R_{\alpha} - R_{\alpha'}| \leq s - 1$ w.p. 1. This establishes 1; 2 follows similarly. \square

We note that, unfortunately, the bounds given in Theorem 7.1 are very weak. This is due to the fact that, for any values of α and α' , we expect $\text{cor}(R_{\alpha}, R_{\alpha'})$ to be positive. Indeed, it has been noted (see, e.g., [3]) that all good criteria are typically highly positively correlated, and our combined criteria would of course be no exception to this general rule. A large positive correlation between the criteria would greatly reduce the variances computed in the proof above, and correspondingly give a lower bound for σ_{\max}^2 . However, the correlation between criteria is merely an empirical fact, and is difficult to establish rigorously except in a few special cases. Thus, in practice we expect σ_{\max}^2 to be substantially smaller, but achieving better rigorous bounds seems to be a difficult problem.

8 The role of the Prior Distributions

As is frequently the case when a statistical method involves the use of prior distributions, a difficulty one must resolve is how to choose an appropriate prior. In our case, we have 2 prior distributions to consider, $\pi_{\mathcal{M}}$ and $\pi_{\Theta(\mathcal{M})}$. In the case of $\pi_{\mathcal{M}}$, as mentioned above, often a prior with simple structure is sufficient. For example, one might choose a prior on the *order* of the true model based on the effect sparsity principle, and, conditioning on the order, assume that all models of that order are equally likely. Of course, knowledge about the specific problem should also be incorporated whenever possible.

The prior $\pi_{\Theta(\mathcal{M})}$ is somewhat more difficult to specify in a principled way. First, we must note that the parameters $\Theta(M_i)$ are actually *nuisance* parameters, since we are not actually interested in doing any inference on them. However, it is our empirical finding that the specification of this prior distribution typically has little effect on the *shape* of the resulting response surface, although it may cause a shift depending on the problem. It seems that, at least in this context, knowledge of the particular distribution from which the parameters are assumed to be drawn is not especially important. Nevertheless, we still recommend adopting a common strategy in Bayesian modelling, which is to repeat an experiment with several different prior distributions in order to assess the sensitivity of the problem to the modelling assumptions. Perhaps more interesting, however, is the observation that the response surface seems to be rather sensitive to the magnitude of the parameters, as we observe that the minimum of the surface shifts from the BIC corner to the PRESS corner as we increase the variance of the prior distribution. Thus, it seems that the only truly important aspect of the specification of $\pi_{\Theta(\mathcal{M})}$ is the order of magnitude of the variance, while the particular shape of distribution chosen has little effect. Here, one should rely on domain-level knowledge of the problem under consideration. We also recommend standardizing the predictors to ensure that the estimated regression coefficients always have a common scale.

9 Extensions

There is still much investigation left to be done in this area. In particular, we have only begun to explore the effects of the many factors that may affect the optimal MSC — the size of the data set, the distributions of the parameters, the correlation structure of the covariates, etc. In particular, a characterization of situations in which combined criteria outperform traditional criteria would be a very useful advance.

Our method, as described, is entirely empirical. In general, we expect analytical results will be very difficult to derive, due to the complexity of the distributions involved. However, we expect that in certain simple, special cases, analytical results may be feasible. It would be interesting to compare the empirical simulation results from our method with these mathematical results. One example of a simple special case would be regression in which the predictor matrix is constrained to be orthogonal.

Further, there are many other MSCs which may be incorporated into the convex combinations. We chose the ones used here mainly due to their readily available software implementations, and the desire to limit the number of base MSCs to 3 in order to make the results easy to visualize. Nevertheless, there is no a priori reason to exclude or favor certain MSCs over others, as our procedure should place weight only on MSCs which have proven power to discriminate between good and poor models.

It would also be useful if there were more efficient procedures for handling the implicit multiple comparisons issue discussed in Section 7. This would allow us to decide whether a result is statistically significant using less computation. We anticipate that the key lies in exploiting the fact that all good MSCs, and combinations thereof, are typically highly correlated with each other. Taking advantage of this fact would likely reduce the Type I error rate in the hypothesis

tests.

References

- [1] Mark H. Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.
- [2] Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-De-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2001.
- [3] Roland T. Rust, Duncan Simester, Roderick J. Brodie, and V. Nilikant. Model selection criteria: an investigation of relative accuracy, posterior probabilities, and combination of criteria. *Management Science*, 41(2):322–333, 1995.