The Canadian Journal of Statistics
Vol. 32, No. 3, 2004, Pages 269–283
La revue canadienne de statistique

Improvement over Bayes prediction in small samples in the presence of model uncertainty

Hubert WONG and Bertrand CLARKE

Key words and phrases: Forecasting; model averaging; mongrel risk.

MSC 2000: Primary 62M20; secondary 62A01.

Abstract: In an online prediction context, the authors introduce a new class of mongrel criteria that allow for the weighing of candidate models and the combination of their predictions based both on model-based and empirical measures of their performance. They present simulation results which show that model averaging using the mongrel-derived weights leads, in small samples, to predictions that are more accurate than that obtained by Bayesian weight updating, provided that none of the candidate models is too distant from the data generator.

Amélioration de la prévision bayésienne dans les petits échantillons en présence d'incertitude à propos du modèle

Résumé : Dans un contexte de prévision continue, les auteurs proposent une nouvelle classe de critères "métissés" permettant de pondérer différents modèles envisagés et de combiner leurs prévisions à partir de mesures fondées sur ces modèles et sur leur performance empirique. Ils font état de simulations montrant que la synthèse de modèles au moyen de poids métissés conduit, dans de petits échantillons, à des prévisions plus précises que celle obtenue par mise à jour bayésienne des poids, pourvu qu'aucun des modèles en cause ne soit trop éloigné de celui dont émanent les données.

1. INTRODUCTION

Suppose we want to predict outcomes Y_n sequentially for $n = 1, 2, \ldots$ based on a vector of explanatory variables with outcomes X_n. Denote the sequence of predictions by $\widehat{Y}_n$, where each $\widehat{Y}_n$ is a function of the X_t with $t \leq n$ and the Y_t with $t \leq n-1$. In regression settings, the differences $Y_n - \widehat{Y}_n$ are ancillary to the unknown parameters in the model. Although ancillary statistics give no information about the parameters directly, the differences are informative about the predictive accuracy of the forecasting procedure as a whole.

In most applications, there are multiple candidate models for the data since the true data generator is unknown. Each model yields a different sequence of predictors, and hence we might construct a new predictor by averaging the predictors from these models at each time point. In Bayesian model averaging (BMA), the weights used in the averaging are given by the posterior probabilities (conditional on the full data) of the models. As an alternative, one could base the weights on the past predictive accuracy of the models. In this paper, we present a "mongrel" forecasting procedure in which the weights are obtained by conditioning on some of the differences $Y_n - \widehat{Y}_n$ from the candidate models, not on the full data. (The term *mongrel* reflects our use of a mixture of model based and empirical criteria which we describe below.) We compare the performance of the mongrel procedure to BMA in simulations and show that roughly, the mongrel procedure outperforms BMA in small sample sizes so long as the candidate models are not too far from the data generator.

The example that motivated our inquiry was the following. Suppose data are generated from the model $Y_t = \gamma_0 + \gamma_1 X_{1,t} + \gamma_2 X_{2,t} + \varepsilon_t$, $t = 1, \ldots, n$, where $\gamma_0 = 1$, $\gamma_1 = 0.8$, $\gamma_2 = 0$, and the ε_t are independent standard normal errors. We do not know which X variables are useful, so we fit two models: a full model which estimates $(\gamma_0, \gamma_1, \gamma_2)$, and a reduced model which estimates (γ_0, γ_1) only. For simplicity, assume the priors on the coefficients in the candidate models are normal with the correct means and identity variance matrix, and assume equal prior

weights on both models. Note that because $\gamma_2 = 0$, the reduced model generally should yield better predictors since both models are unbiased for the parameters but the reduced model has less estimation error.

We evaluated the predictive performance of BMA and of a naive version of the mongrel procedure that computes posterior weights based on only the last $n/2$ differences from the smaller model. The results of this simple comparison are shown in Figure 1. We used $m = 5000$ sequences of length 40 and evaluated the mean squared prediction error (MSPE)

$$\text{MSPE} = \frac{1}{m}\sum_{i=1}^{m}(Y_{n+1} - \widehat{Y}_{n+1})^2$$

for each time step $n = 1, \ldots, 40$. The solid line in the top panel of Figure 1 is the mean squared prediction error from BMA. Below it, the dashed line, is the mean squared prediction error from the naive mongrel approach (labeled "$n/2$"). The middle panel in Figure 1 shows that the difference in mean squared prediction error, with error bars, is systematically positive and clearly favours the mongrel approach. More importantly, this reduction in mean squared prediction error occurred despite the fact that BMA gave higher weight to the reduced model on average than the mongrel approach (bottom panel of Figure 1)! This counter-intuitive result represents compelling evidence that the differences $Y_t - \widehat{Y}_t$ can be more informative than all of the data when comparing the predictive performance of different models. Unfortunately, the improvement obtained here does not hold when γ_2 is large. Conditioning on a preselected number of the most recent differences is too coarse a strategy; the mongrel procedure we define chooses adaptively how many differences to condition on and appears to outperform BMA across a wide range of scenarios.

Our comparisons here are based on predictive error. We regard this criterion as fundamental because it satisfies the "prequential principle" (Dawid 1984) in that it evaluates predictors independently of their method of construction. This means that all predictors, regardless of their origin, compete according to a uniform standard, set only by the data. The criterion does not favour any one method over another.

Existing criteria that have been used for constructing the weights fall into one of two classes that we call "model-based" and "empirical." Model-based criteria depend on an assumed probability model. For instance, a likelihood or an expected risk of a predictor computed conditionally on the model and the data is a model-based criterion. In general, two different models will generate different values for the criterion even if both models give the same sequence of predictions in the past. The weights used in BMA are model-based since posterior probabilities are obtained from the marginal densities of the candidate models. Recent reviews include Raftery, Madigan & Hoeting (1997), Clyde (1999), and Hoeting, Madigan, Raftery & Volinsky (1999). Note that posterior probabilities reflect the fit of the data to the model rather than evaluating the expected accuracy of the current prediction.

In contrast, an empirical criterion assesses the worth of a model strictly on its observed predictive performance. For instance, the worth of the predictor $\widehat{Y}_{k,t}$ from model k could be the loss $L(\widehat{Y}_{k,t}, Y_t)$ evaluated on the observed values only. In particular, if two models give the same predictions they have the same sequence of losses and are judged equally good without regard to the structure of the underlying models. (Indeed, an empirical criterion does not even require that the predictions derive from a model; all that is required is that the forecasting procedure issues a prediction at each time point.) The paradigmatic empirical criterion is "leave-one-out" cross-validation; see Mosteller & Tukey (1968). If the Y_j are omitted one at a time and $\widehat{Y}_{k,j}$ is obtained by fitting the model with the remaining data, then the loss for using model k is $\sum_j L(\widehat{Y}_{k,j}, Y_j)$. When L is a squared error loss, this gives the predicted residual sum of squares, or PRESS (Allen 1974). In a sequential setting, however, the PRESS criterion is artificial because the prediction for a given time point should not be constructed using data from later time points. Dawid (1984) corrected this by suggesting $L(\widehat{Y}_{k,n}, Y_n)$, $n = 1, \ldots$, the sequence of one-

step ahead prediction losses, as a basis for evaluating models. Here, $\widehat{Y}_{k,n}$ is based only on the data available at time point $n-1$. Subsequent work by Dawid (1992), Sellier-Moiseiwitsch & Dawid (1993), Skouras & Dawid (1999) established consistency and efficiency for some of these prequential procedures.

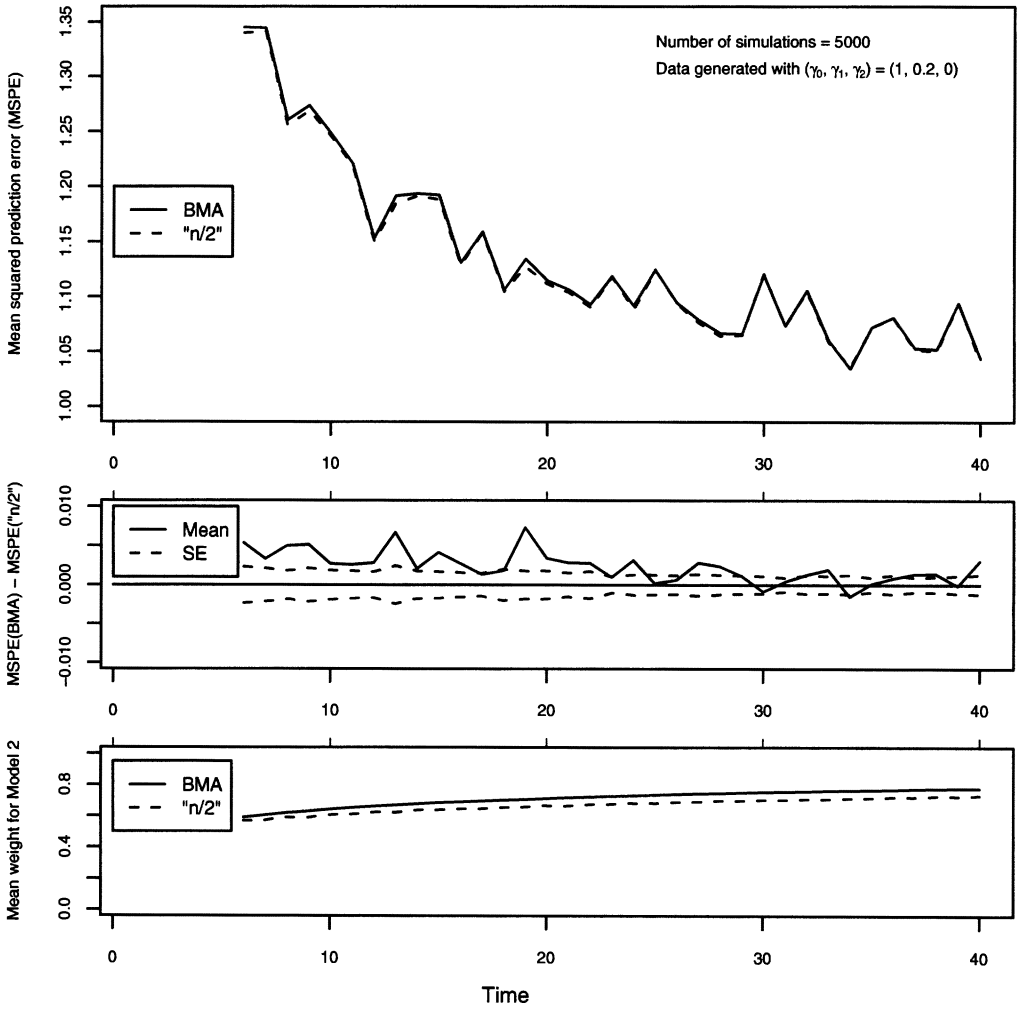

FIGURE 1: Performance of Bayes model averaging (BMA) and a naive mongrel prediction approach ("$n/2$"). *Top panel:* Mean squared prediction error (MSPE) for BMA (solid line) and the "$n/2$" approach (dashed line). *Middle panel:* MSPE(BMA) − MSPE("$n/2$") (solid line) with standard error for the difference (dashed line). *Bottom panel:* Average weight assigned to the reduced model (BMA - solid curve, "$n/2$" - dashed curve).

Our mongrel procedure combines aspects of both the model-based and the empirical approaches. We retain a probabilistic framework for evaluating the adequacy of each model. Rather than on full data, however, the evaluations condition on statistics $\mathbf{S}_n$ that reflect the past predictive performance of the models. These evaluations result in weights that are functions of $\mathbf{S}_n$ rather than of the full data. We formally describe the mongrel procedure in Section 2.

Section 3 gives the formulae for the implementation of the mongrel procedure in a normal linear regression setting. Since we use them later in our computational work, we focus on sets $\mathbf{S}_n$ consisting of "predictuals," i.e., sets of residuals that would arise from using the predictions from the kth model $Y_t - \widehat{Y}_{k,t}$.

In Section 4, we present our computational work based on normal linear models. The mongrel procedure gives better predictions than the Bayes procedure does, across a range of choices for data generator and model prior. Although our procedure generalizes to any finite number of models, we see that it can be made to break down when some model is far enough from the data generator.

In Section 5, we discuss the implications for methodology and challenges to the concept of model uncertainty and model list selection that our method reveals.

The Appendix gives the derivations of the formulae presented in Section 3.

2. MONGREL RISK

Let $\mathbf{Y} = (Y_1, Y_2, \ldots)$ denote the sequence of random variables that is to be predicted. At each time point n, we must issue a prediction for the value of Y_{n+1}. We assume that the following information is available:

1. A p-vector of covariates $\mathbf{X}_{n+1}$ whose elements may be related to Y_{n+1};
2. The outcomes and covariates already observed up to time point n, i.e., $\mathbf{Y}_{(n)} = (Y_1, \ldots, Y_n)$ and $\mathbf{X}_{(n)}$, the $n \times p$ matrix with row i equal to $\mathbf{X}_i$;
3. A list of candidate models $\mathcal{M} = \{k : k = 1, \ldots, K\}$ wherein each model k describes the structure of the probabilistic dependence of the outcomes on the covariates and a vector of unknown parameters θ_k; and
4. A prior density $\pi_k(\theta_k)$ on each θ_k and a vector of prior probabilities $\alpha_0 = (\alpha_{1,0}, \ldots, \alpha_{K,0})$ on the models.

Model k is true if it contains the true distribution of $\mathbf{Y}$, i.e., the data generator, and no submodel (in $\mathcal{M}$) of model k contains this distribution.

Under squared error loss, the Bayes predictor $\widehat{Y}_{k,n+1}$, conditional on model k being true, is simply the posterior mean of Y_{n+1} under model k, i.e.,

$$\widehat{Y}_{k,n+1} = \mathrm{E}_{k;\mathbf{Y}_{(n)}} Y_{n+1}. \tag{1}$$

Throughout this paper we will use $\mathrm{E}_{k;\mathbf{S}}$ to denote the expectation operator in which model k is assumed to be true and the marginalization occurs over θ_k and the randomness in $\mathbf{Y}_{(n)}$ that is not part of the statistic $\mathbf{S}$. Similarly, $\mathbf{V}_{k;\mathbf{S}}$ and $\mathbf{C}_{k;\mathbf{S}}$ will denote the corresponding variance and covariance operators. An operator with only the subscript k (e.g., E_k) will denote marginalization over θ_k and $\mathbf{Y}_{(n)}$.

Each model gives a forecast and we use these to produce a single forecast for actual use. Thus, we can choose one of the forecasts, in effect choosing the model which produced it—the model choice approach; or we can use a forecast obtained by weighting the forecasts from the models—a model averaging approach.

For model averaging, we must assign a weight to each model. Starting with the prior weights α_0 and given the information contained in a (vector of) statistic(s) $\mathbf{S}_n = \mathbf{S}_n(\mathbf{Y}_{(n)})$, we can apply the Bayes theorem to update α_0 and obtain the posterior weights $\alpha(\mathbf{S}_n) = (\alpha_1(\mathbf{S}_n), \ldots, \alpha_K(\mathbf{S}_n))$, that is, for each model k,

$$\alpha_k(\mathbf{S}_n) = \frac{\alpha_{k,0} m_k(\mathbf{S}_n)}{\sum_{i=1}^{K} \alpha_{i,0} m_i(\mathbf{S}_n)}, \tag{2}$$

where $m_k(\mathbf{S}_n)$ is the density of $\mathbf{S}_n$ (marginalized over π_k and the randomness in $\mathbf{Y}_{(n)}$ that is not part of $\mathbf{S}_n$) under model k. These posterior weights generate a class of predictors

$$\widehat{Y}_{n+1}(\mathbf{S}_n) = \sum_{k=1}^{K} \alpha_k(\mathbf{S}_n) \widehat{Y}_{k,n+1}$$

indexed by the choice of $\mathbf{S}_n$. (The dependence of $\widehat{Y}_{k,n+1}$ on $\mathbf{Y}_{(n)}$ has been suppressed to simplify the notation.) Thus, in our approach, choosing the predictor for Y_{n+1} is equivalent to choosing the $\mathbf{S}_n$ used to compute the posterior weights. When $\mathbf{S}_n = \mathbf{Y}_{(n)}$, these posterior weights simply generate the predictors associated with a pure Bayes approach, i.e., with BMA.

Our main point is that by choosing $\mathbf{S}_n$ to be a statistic that reflects the past empirical performance of the models, we often obtain more accurate predictions than by always using $\mathbf{S}_n = \mathbf{Y}_{(n)}$. Natural quantities to include in $\mathbf{S}_n$ are the squared-error losses $(Y_t - \widehat{Y}_{k,t})^2$ from previous time points t that would have been incurred had the predictor $\widehat{Y}_{k,t}$ been used. As alternatives, we use the following quantities.

DEFINITION 1. The *predictual* arising from predicting Y_t using the predictor $\widehat{Y}_{k,t}$ as defined in (1) is

$$R_{k,t} = Y_t - \widehat{Y}_{k,t}.$$

By conditioning on a statistic $\mathbf{S}_n$ that includes losses or predictuals, we obtain predictors that are functions of the actual performance of the models rather than simply on data values. We will focus on statistics $\mathbf{S}_n$ that include predictuals rather than losses for two reasons. The conceptual reason is that losses do not distinguish between the bias and variance components in the error, and this information may be relevant to assessing the quality of the candidate models. The pragmatic reason is that in normal linear models, using predictuals (or any affine functions of $\mathbf{Y}_{(n)}$) lets us easily evaluate quantities such as the posterior weights and others we will introduce later.

Our primary goal is to find good choices for $\mathbf{S}_n$. Observe that as t increases, the differences $Y_t - \widehat{Y}_{k,t}$ are based on ever more data. Since the variance of $Y_t - \widehat{Y}_{k,t}$ decreases as t increases, we suspect that $Y_1 - \widehat{Y}_{k,1}$ is less informative (i.e., a poorer indicator of the predictive accuracy of a model) than later values of $Y_t - \widehat{Y}_{k,t}$. Also, note that the information content of the vector of differences $Y_t - \widehat{Y}_{k,t}$ for $t = 1, \ldots, n$ is equivalent to the information content of the full data set because they are mathematically equivalent, i.e., they generate the same σ-field. So, even though the information gain per time step is the same, the information per difference is higher for differences that appear later and the information in a set of statistics increases as the set shifts forward in time. For instance, the information in $Y_t - \widehat{Y}_{k,t}$ with $t = n$ is higher on average than the information in any other difference with $t < n$. Therefore, we will only consider collections of predictuals for which the inclusion of a predictual from time point t means that all later predictuals are also included. The special case of using all of the predictuals from past time points for any model is equivalent to using the Bayes procedure since the σ-field generated by all predictuals equals the σ-field generated by the data.

Indeed, any set of n linearly independent predictuals will replicate the effect of conditioning on the whole data set. As an empirical fact, computations not described here show that conditioning on a set of predictuals of dimension much less than the sample size n can reproduce the effect of conditioning on the whole data set. For instance, in the simulation framework considered in Section 4 with two nested models differing by a single predictor, the most recent predictual from both models together reproduces the effect of conditioning on all the data. To avoid this kind of undesirable reduction, we restrict choices for $\mathbf{S}_n$ to include predictuals from one of the models only.

A naive specification of $\mathbf{S}_n$ would include only the $n/2$ most recent predictuals, an example which was described in the introduction. However, this specification performs well only for a limited variety of scenarios. A better approach is to choose at each time point the form of $\mathbf{S}_n$ by an optimality criterion that is adaptive to the preceding data sequence. Here, we evaluate a novel type of risk for each candidate $\mathbf{S}_n$ and select the $\mathbf{S}_n$ that minimizes this risk. As the true model is unknown, we first consider the risk of using $\mathbf{S}_n$ under each model.

DEFINITION 2. The *mongrel risk* of the predictor $\widehat{Y}_{n+1}(\mathbf{S}_n)$ assuming model k is true is

$$\rho_k\{\widehat{Y}_{n+1}(\mathbf{S}_n)\} = \mathrm{E}_{k;\mathbf{S}_n}(Y_{n+1} - \widehat{Y}_{n+1})^2. \tag{3}$$

The distinguishing feature of this risk is that the expectation is conditional on $\mathbf{S}_n$, not on the full data. Let $\mathcal{S}_n$ denote the collection of $\mathbf{S}_n$ under consideration at time point n.

DEFINITION 3. The *risk profile* when model k is true is the collection of mongrel risks $\rho_k\{\widehat{Y}_{n+1}(\mathbf{S}_n)\}$ generated by varying $\mathbf{S}_n$ over $\mathcal{S}_n$.

Examination of the risk profiles for all of the models allows us to compare the relative adequacy of the different choices for $\mathbf{S}_n$ both within and across the models. Ideally, the chosen $\mathbf{S}_n$ should result in a low risk for $\widehat{Y}_{n+1}(\mathbf{S}_n)$ regardless of which model is true. Hence one optimality criterion for selecting $\mathbf{S}_n$ is to find for each $\mathbf{S}_n$ the maximum mongrel risk over all of the models and then choose the $\mathbf{S}_n$ that minimizes the maximal risk. This leads to the following predictor.

DEFINITION 4. At each time point n, the *mongrel predictor for* Y_{n+1} is $\widehat{Y}_{n+1}(\mathbf{S}_n^*)$, where

$$\mathbf{S}_n^* = \arg \min_{\mathbf{S}_n \in \mathcal{S}_n} \max_k \rho_k\{\widehat{Y}_{n+1}(\mathbf{S}_n)\}. \tag{4}$$

We have presented here the mongrel procedure in its simplest form. Possible generalizations and alternate formulations include: (i) the optimality criterion could invoke minimization of some average risk over models for each $\mathbf{S}_n$ rather than minimization of the maximum risk; (ii) the expectation in (3) could be evaluated conditional on a statistic different from $\mathbf{S}_n$; and (iii) the specification that we use the Bayes predictors (given by (1)) as the predictors to be averaged could be relaxed, i.e., we could use instead mongrel-type predictors from each model obtained by minimizing expected loss conditional on some statistic rather than on all of the data. See Wong (2000) for a more general development.

3. FORMULAE FOR NORMAL LINEAR MODELS

Implementing the mongrel procedure requires evaluating the posterior weights in (2) and the mongrel risks in (3). For general classes of candidate models and forms of $\mathbf{S}_n$, these evaluations typically will be difficult to perform. However, when the candidate models are from the class of normal linear models, the loss function is a squared error, and $\mathbf{S}_n$ is affine in $\mathbf{Y}_{(n)}$, then analytic formulae can be derived. We give these formulae in this section.

Consider the collection of subset linear regression models

$$\mathbf{Y}_{(n)} \mid \mathbf{X}_{(n)}, \beta_k \sim \mathsf{N}\big(\mathbf{X}_{(n)}\mathbf{D}_k\beta_k, \sigma^2\mathbf{I}\big)$$

indexed by k, where $\mathbf{D}_k$ is a $p \times p_k$ matrix of zeros and ones that picks out the p_k covariates in model k (Allen 1974) and $\mathsf{N}(a, b)$ denotes the normal distribution with mean vector a and variance matrix b. For simplicity, assume that σ^2 is known and equip the parameter vector β_k with the prior distribution π_k given by

$$\beta_k \sim \mathsf{N}(\mathbf{b}_k, \Gamma_k).$$

The marginal distribution for $\mathbf{Y}_{(n)}$ under model k after mixing over π_k is $\mathsf{N}(\nu_{k,n}, \Psi_{k,n})$ where

$$\nu_{k,n} = \mathbf{Z}_{k,(n)}\mathbf{b}_k, \quad \Psi_{k,n} = \sigma^2\mathbf{I} + \mathbf{Z}_{k,(n)}\Gamma_k\mathbf{Z}_{k,(n)}^\top$$

and $\mathbf{Z}_{k,(n)} = \mathbf{X}_{(n)}\mathbf{D}_k$.

Suppose $\mathbf{S}_n$ is a $J = J(n)$ vector expressed in the form

$$\mathbf{S}_n = \mathbf{U}^\top(\mathbf{Y}_{(n)} + \mathbf{c}), \tag{5}$$

where the $n \times J$ matrix $\mathbf{U}$ and the n-vector $\mathbf{c}$ do not depend on $\mathbf{Y}_{(n)}$. Without loss of generality, we assume $\mathbf{U}$ to be of full rank. In fact, we are interested in $\mathbf{S}_n$ only for the σ-field it generates. (The constant $\mathbf{c}$ is included to ensure the class $\mathbf{S}_n$ includes predictuals.)

From the properties of the multivariate normal distribution and the fact that $\mathbf{S}_n$ is affine in $\mathbf{Y}_{(n)}$, if model i is true, then $\mathbf{S}_n \sim \mathsf{N}(\mu_i, \Sigma_i)$ with

$$\mu_i = \mathbf{U}^\top \mathbf{Z}_{i,(n)} \mathbf{b}_i + \mathbf{U}^\top \mathbf{c}, \quad \Sigma_i = \sigma^2 \mathbf{U}^\top \mathbf{U} + \mathbf{U}^\top \mathbf{Z}_{i,(n)} \Gamma_i \mathbf{Z}_{i,(n)}^\top \mathbf{U}.$$

This result immediately gives the posterior weights through setting

$$m_i(\mathbf{S}_n) = (2\pi)^{-J/2} |\Sigma_i|^{-1/2} \exp\{-\tfrac{1}{2} (\mathbf{S}_n - \mu_i)^\top \Sigma_i^{-1} (\mathbf{S}_n - \mu_i)\}$$

as the marginal density used to evaluate (2).

The mongrel risks in (3) can be computed using

$$\begin{aligned} \rho_i\{\widehat{Y}_{n+1}(\mathbf{S}_n)\} &= \sum_{j=1}^{K} \sum_{k=1}^{K} \alpha_j(\mathbf{S}_n) \alpha_k(\mathbf{S}_n) \{\mathbf{C}_i(R_{j,n+1}, R_{k,n+1}) - \Xi_{j,i} \Sigma_i^{-1} \Xi_{k,i}^\top\} \\ &\quad + \left[\sum_{k=1}^{K} \alpha_k(\mathbf{S}_n) \{\mathrm{E}_i R_{k,n+1} + \Xi_{k,i} \Sigma_i^{-1} (\mathbf{S}_n - \mu_i)\}\right]^2, \end{aligned} \tag{6}$$

where

$$\Xi_{k,i} = \left(\Psi_{i,n}^{-1} \mathbf{Z}_{i,(n)} \Gamma_i \mathbf{Z}_{i,n+1} - \Psi_{k,n}^{-1} \mathbf{Z}_{k,(n)} \Gamma_k \mathbf{Z}_{k,n+1}\right)^\top \Psi_{i,n} \mathbf{U} \tag{7}$$

is the covariance between $R_{k,n+1}$ and $\mathbf{S}_n$ under model i,

$$\mathrm{E}_i R_{k,n+1} = \mathbf{u}_{k,n+1}^\top \left(\mathbf{Z}_{i,(n+1)} \mathbf{b}_i - \mathbf{Z}_{k,(n+1)} \mathbf{b}_k\right), \tag{8}$$

$$\mathbf{C}_i(R_{j,n+1}, R_{k,n+1}) = \mathbf{u}_{j,n+1}^\top \Psi_{i,n+1} \mathbf{u}_{k,n+1} \tag{9}$$

and

$$\mathbf{u}_{k,n+1}^\top = \left(-\mathbf{Z}_{k,n+1}^\top \Gamma_k \mathbf{Z}_{k,(n)}^\top \Psi_{k,n}^{-1}, 1\right). \tag{10}$$

See the Appendix for the derivations of (6), (8), and (9).

All of the preceding formulae in this section apply to $\mathbf{S}_n$ that are arbitrary affine functions of $\mathbf{Y}_{(n)}$, i.e., for arbitrary (compatibly dimensioned) specifications of $\mathbf{U}$ and $\mathbf{c}$ in (5). When $\mathbf{S}_n$ consists of predictuals, the computational burden can be reduced because $\mathbf{U}$ and $\mathbf{c}$ can be constructed using quantities already presented. From inspection of (13) in the Appendix, we see that to include $R_{k,t}$ (for any $t \leq n$) in $\mathbf{U}$, we simply set (i) the first t elements in the row in $\mathbf{U}$ corresponding to $R_{k,t}$ to $\mathbf{u}_{k,t}^\top$ and the remaining elements to zeros, and (ii) $\mathbf{c} = -\mathbf{Z}_{k,(n)} \mathbf{b}_k$.

The solution to the optimization used to find $\mathbf{S}_n^*$ cannot be expressed in closed form. This does not pose a difficulty in finding the mongrel predictor, as the solution is determined easily by searching over the risk profiles. However, the absence of a simple analytic solution means that the performance of the mongrel predictor needs to be evaluated by simulation.

4. SIMULATION STUDY

4.1. Simulation framework.

We used the following simulation framework to assess the forecasting performance for different methods for specifying $\mathbf{S}_n$. Data sequences of length 40 were generated randomly from the model

$$Y_n = \gamma_0 + \gamma_1 X_{1,n} + \gamma_2 X_{2,n} + \varepsilon_n,$$

where $X_{1,n}$, $X_{2,n}$, ε_n were all independent standard normal variables. For the base set of scenarios, we fixed $\gamma_0 = 1$ and $\gamma_1 = 0.2$, while γ_2 was varied over the set $\{0, 0.2, 0.4\}$.

To assess the performance of the mongrel procedure when the collection of candidate models consists of two nested models only, we considered:

- Model 1 (the "Full" model): containing the intercept and both X_1 and X_2, i.e.,

$$Y_n = \beta_0 + \beta_1 X_{1,n} + \beta_2 X_{2,n} + \varepsilon_n;$$

- Model 2 (the "X_1-only" model): containing the intercept and X_1 only, i.e.,

$$Y_n = \beta_0^{(1)} + \beta_1^{(1)} X_{1,n} + \varepsilon_n.$$

For prior distributions on the parameters, we assumed that

$$(\beta_0, \beta_1, \beta_2) \sim \mathsf{N}\big((\gamma_0, \gamma_1, \gamma_2), \mathbf{I}\big) \quad \text{and} \quad \big(\beta_0^{(1)}, \beta_1^{(1)}\big) \sim \mathsf{N}\big((\gamma_0, \gamma_1), \mathbf{I}\big).$$

We considered three choices for α_0, the prior probability on the models: $(1/4, 3/4)$, $(1/2, 1/2)$, $(3/4, 1/4)$. Hence the base set included nine scenarios (three choices for γ_2 times three choices for α_0). The collection of $\mathbf{S}_n$ over which we optimized was

$$\mathcal{S}_n^2 = \big\{\mathbf{R}_{2,J} : J \in 1, \ldots, n-1\big\},$$

where $\mathbf{R}_{2,J} = (R_{2,n-1}, \ldots, R_{2,n-J})$, i.e., the most recent J predictuals from Model 2. The number of sequences used in each scenario was $m = 5000$.

To assess the performance for a nonnested collection of candidate models, we added a third model,

- Model 3 (the "X_2-only" model): containing the intercept and X_2 only, i.e.,

$$Y_n = \beta_0^{(2)} + \beta_1^{(2)} X_{2,n} + \varepsilon_n$$

with the prior distribution

$$\big(\beta_0^{(2)}, \beta_1^{(2)}\big) \sim \mathsf{N}\big((\gamma_0, \gamma_2), \mathbf{I}\big).$$

Again, we considered three choices for α_0: $(1/3, 1/3, 1/3)$, $(1/2, 1/6, 1/3)$, $(1/6, 1/2, 1/3)$. As before, this results in a set of nine scenarios. Here, the collection of $\mathbf{S}_n$ over which we optimized was

$$\mathcal{S}_n^3 = \big\{\mathbf{R}_{3,J} : J \in 1, \ldots, n-1\big\},$$

where $\mathbf{R}_{3,J} = (R_{3,n-1}, \ldots, R_{3,n-J})$. The use of $\mathcal{S}_n^3$ rather than $\mathcal{S}_n^2$ is somewhat arbitrary and both choices yield similar performance results.

Note that in the above scenarios, all of the candidate models are "close" to the data generator (irrespective of the value used for γ_2). In fact, the greatest separation occurs when $\gamma_2 = 0.4$ and Model 2 is used, i.e., the worst model corresponds to one that omits a predictor that (conditional on the remaining terms) has a correlation of 0.37 with the outcome. (With unit variances for $X_{2,n}$ and ε_n, the correlation between Y and X_2, conditional on X_1, is $\gamma_2/\sqrt{1+\gamma_2^2}$, which evaluates to 0.37 when $\gamma_2 = 0.4$.)

To investigate the sensitivity of the mongrel procedure to including a distant model in the set of candidate models, we repeated the evaluation of the above scenarios but with $\gamma_1 = 0.8$. With this specification, Models 1 and 2 remain close to the data generator as before, but Model 3 is relatively distant as it omits a predictor that has a correlation of 0.60 with the outcome. Failing to include an obviously important variable is an unusual choice for a candidate model. However, we have done so because in practice this unfortunate situation does occur.

Additional simulations were performed to assess the sensitivity to the priors on the regression coefficients, in particular to the use of diffuse priors and to misspecified prior means.

The performance of the mongrel procedure and of BMA were compared using the empirical mean squared prediction error

$$\text{MSPE} = \frac{1}{m}\sum_{i=1}^{m}(Y_{n+1} - \widehat{Y}_{n+1})^2.$$

All computations were performed in the "R" statistical programming environment.

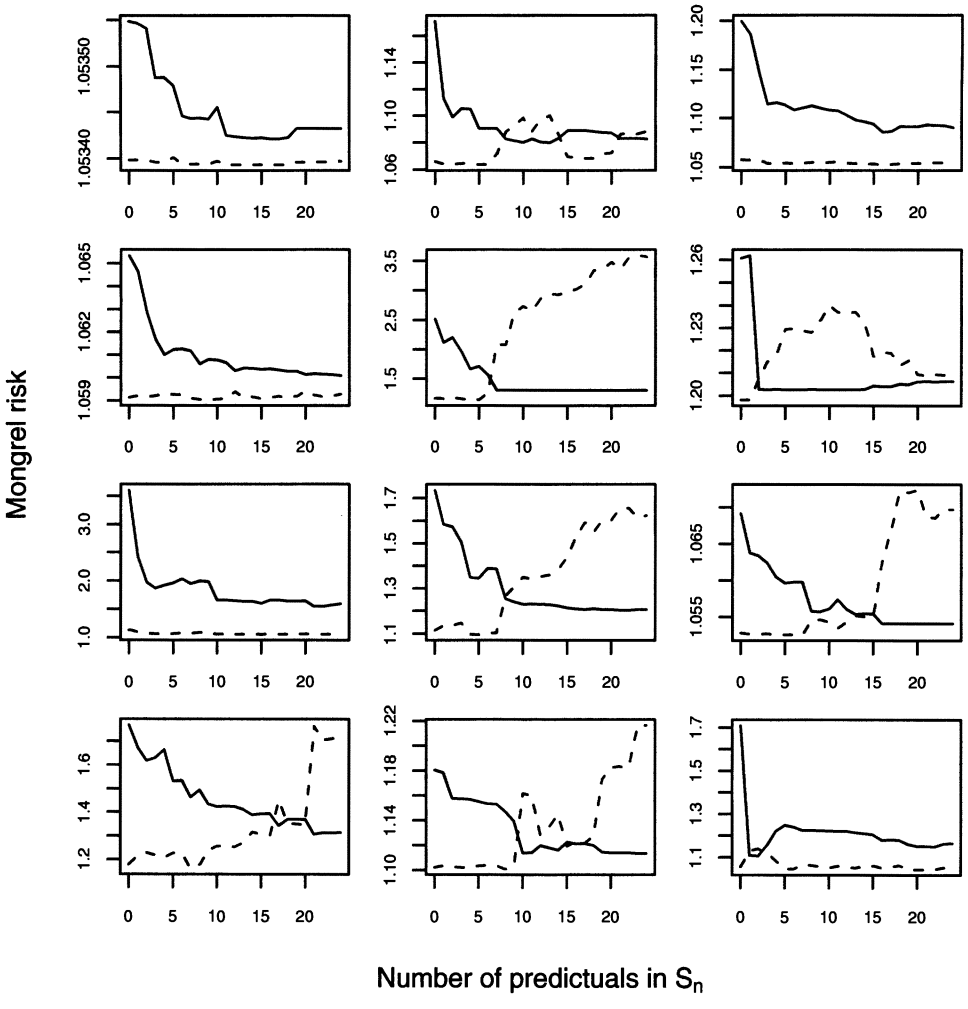

FIGURE 2: Mongrel risk profiles for the first 12 sequences. The solid line and dashed lines assume that the "Full" and "X_1 only" models are true, respectively.

4.2. Results for the 2-model case.

To illustrate the key features of the mongrel procedure, we present in detail the case in which $\gamma_1 = 0.2$ and the model list contains only Models 1 and 2. Figure 2 gives examples of risk profiles, i.e., of how the mongrel risk varies as a function of the number of predictuals included in $\mathbf{S}_n$ for predicting the outcome at time point 25. Individual panels correspond to each of the first 12 sequences taken from the simulations with $\gamma_2 = 0.4$. The solid line obtains when Model 1 is assumed to be true and the dashed line obtains when Model 2 is assumed to be true. Note that under Model 1, the mongrel risk tends to decrease as the number of predictuals increases. This pattern reflects the intuition that the parameters tend to be more accurately estimated with additional information leading to more accurate predictions. However, the decrease is not monotonic because of stochasticity; at times, some predictuals are not merely uninformative, but are in fact misleading. Under Model 2, the mongrel risk sometimes increases as the number of predictuals increases because the additional information identifies the prediction error due to bias (through omitting X_2) in the model. In the scenarios with $\gamma_2 = 0$, this pattern occurred much less frequently since then Model 2 is unbiased (graphs not shown).

The only sequence that devolved to choosing the full data (i.e., 24 predictuals) was the one in panel [2, 1]. Approximately 15% of the 5000 sequences selected full data as the optimal conditioning statistic and 6% of the sequences selected 23 predictuals (i.e., all except the predictual from the first time point). In general, the frequency of selecting a given $\mathbf{S}_n$ decreased as the number of predictuals in $\mathbf{S}_n$ decreased. Less than 1% of the sequences selected only a single predictual as the optimal conditioning statistic. When $\gamma_2 = 0$ or 0.2, the full data were selected with somewhat greater frequency but the trends are similar.

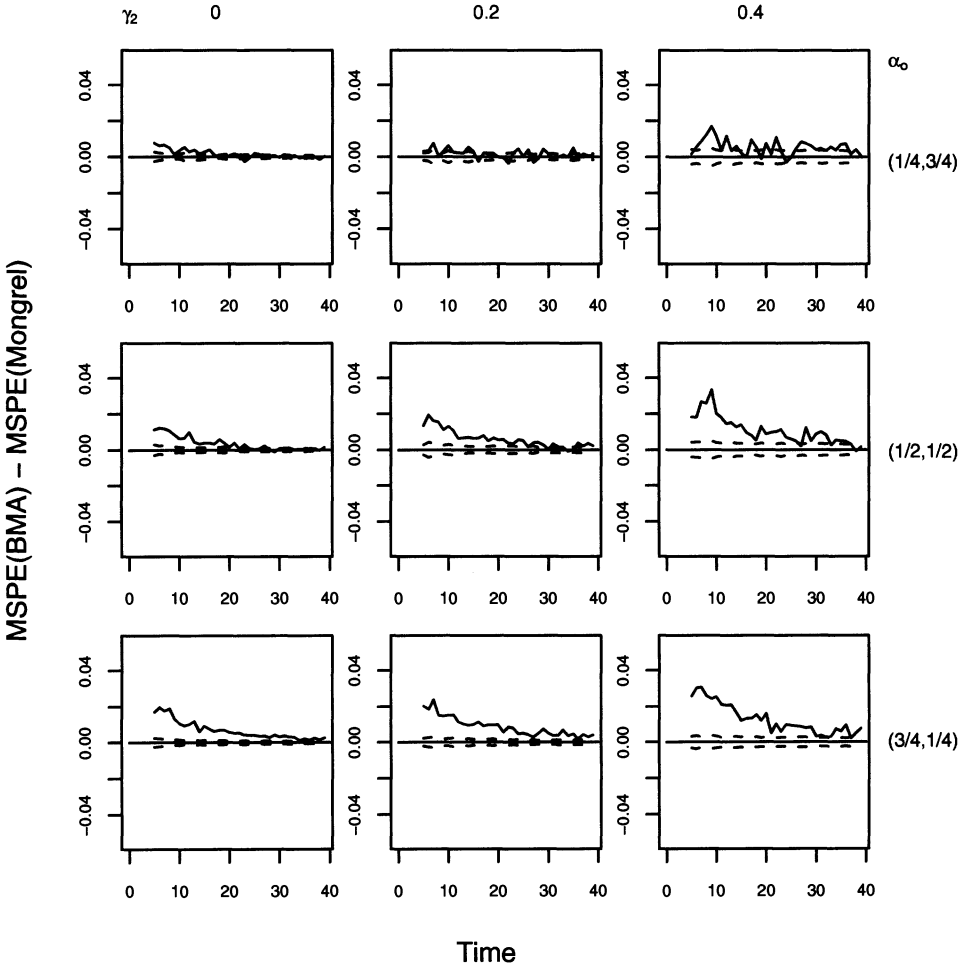

FIGURE 3: Performance of the mongrel procedure relative to BMA for different values of γ_2 and different prior model weights α_o when $\gamma_1 = 0.2$ and the model list has 2 models. The solid line is MSPE(BMA) − MSPE(Mongrel). The dashed lines represent ± 1 standard error for the difference.

The predictive performance of the mongrel procedure compared to BMA is summarized in Figure 3. The solid line (MSPE(BMA) − MSPE(Mongrel)) is the MSPE arising from using BMA minus the MSPE arising from using the mongrel procedure; values above the horizontal axis favour the mongrel procedure. The dashed lines represent ± 1 standard error of the difference. The three rows of panels correspond to the choices of α_0 and the three columns correspond to the choices of γ_2. In the worst case, with $\gamma_2 = 0$ and $\alpha_0 = (3/4, 1/4)$, the mongrel procedure performs about equally well in comparison with BMA. As γ_2 and/or the prior weight assigned to Model 1 increases, the mongrel procedure improves and beats out BMA, especially with smaller

sample sizes.

The results with $\gamma_1 = 0.8$ are nearly identical to the ones just presented with $\gamma_1 = 0.2$ and are not reproduced here.

Figure 4 displays a modified boxplot of MSPE(BMA) − MSPE(Mongrel) for the individual sequences. Within the range $(-1, 1)$, the differences are split roughly equally between positive and negative deviations. However, nearly all of the large magnitude deviations are positive. Thus, the mongrel procedure seldom performs substantially worse than BMA whereas BMA performs much worse than the mongrel procedure on a meaningful number of the sequences. These results suggest that the mongrel procedure is more robust than BMA to atypical data.

4.3. Results for the 3-model case.

Figure 5 summarizes the comparison of the mongrel procedure to BMA when $\gamma_1 = 0.2$ and the model list contains all three models using the same format as Figure 3. The predictive performance of the mongrel procedure improves upon BMA substantially across all nine scenarios and the gains are greater than in the 2-model case. In addition, the graphs suggest that these gains extend to sample sizes larger than 40 (the maximum we considered).

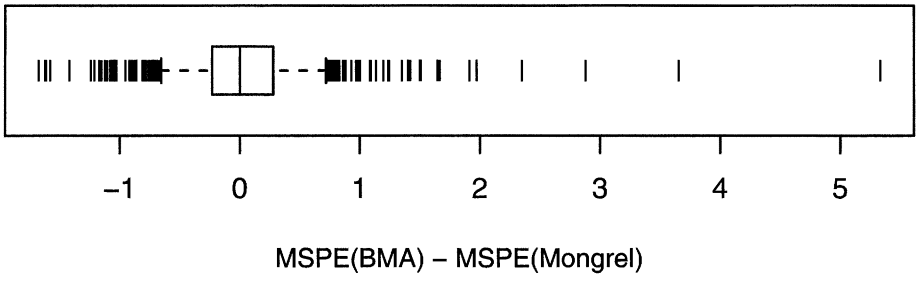

FIGURE 4: Modified boxplot of the differences MSPE(BMA) − MSPE(Mongrel) for the 5000 sequences. The box encompasses the 5th to 95th percentiles, the whiskers extend to the 1st and 99th percentiles. Vertical bars denote values below the 1st or above the 99th percentiles.

When γ_1 is changed from 0.2 to 0.8 and the model list contains all three models, the performance of the mongrel procedure relative to BMA changes dramatically. For very small sample sizes ($n < 10$), the mongrel procedure continued to beat out BMA, but it quickly lost out to BMA as the sample size increased. Moreover, the magnitudes of the difference in performance, both when the mongrel procedure beat or lost out to BMA, typically exceeded those seen in Figure 5, with average differences around 0.05. As before, the performance of the mongrel procedure tended to improve as γ_2 increased. We observed that the mongrel posterior weight for Model 3 converged to zero very slowly in all of these scenarios. Thus, considerable weight was being placed on a very poor model when the Bayes weights essentially had discarded this model.

4.4. Sensitivity to priors on the model parameters.

We performed additional simulations in the 2-model case to assess sensitivity of the procedure to the choice of prior distributions on the model parameters. For our base set, we had set the prior variances, Γ_1 and Γ_2, on the regression parameters to be identity matrices because we felt that such values reflect the typical (small to moderate) amount of prior information available in practice. When very weak priors ($\Gamma_i = 25\mathbf{I}$) were used, the simulations yielded results that were qualitatively the same. Our choice of prior means for the parameters also may seem unduly optimistic in that they are close to or equal the coefficients from the data-generator. In practice, we would expect that the prior means would not match the values in the data-generator, so we should consider other choices. The concern here is that because the mongrel procedure sets aside information, it will be less adept at compensating for "bad" prior means. But what range of priors

is of practical interest? As an arbitrary choice that reflects a reasonably bad prior, we set the prior means to be $(-1, -1, -1)$ and $(-1, -1)$ for Models 1 and 2, respectively. The results were that for the larger sample sizes, now the Bayes procedure slightly beat out the mongrel procedure when $\gamma_2 = 0.4$ and $\alpha_0 = (0.25, 0.75)$ and that the gains obtained using the mongrel procedure in the other scenarios appeared to be attenuated. However, for the smaller sample sizes, the gains obtained using mongrel procedure increased considerably—an odd result given that the initial concern had been that the mongrel procedure would be misled by the bad prior. Overall, the results suggested that the mongrel procedure is only slightly more sensitive to bad priors than is the Bayes procedure.

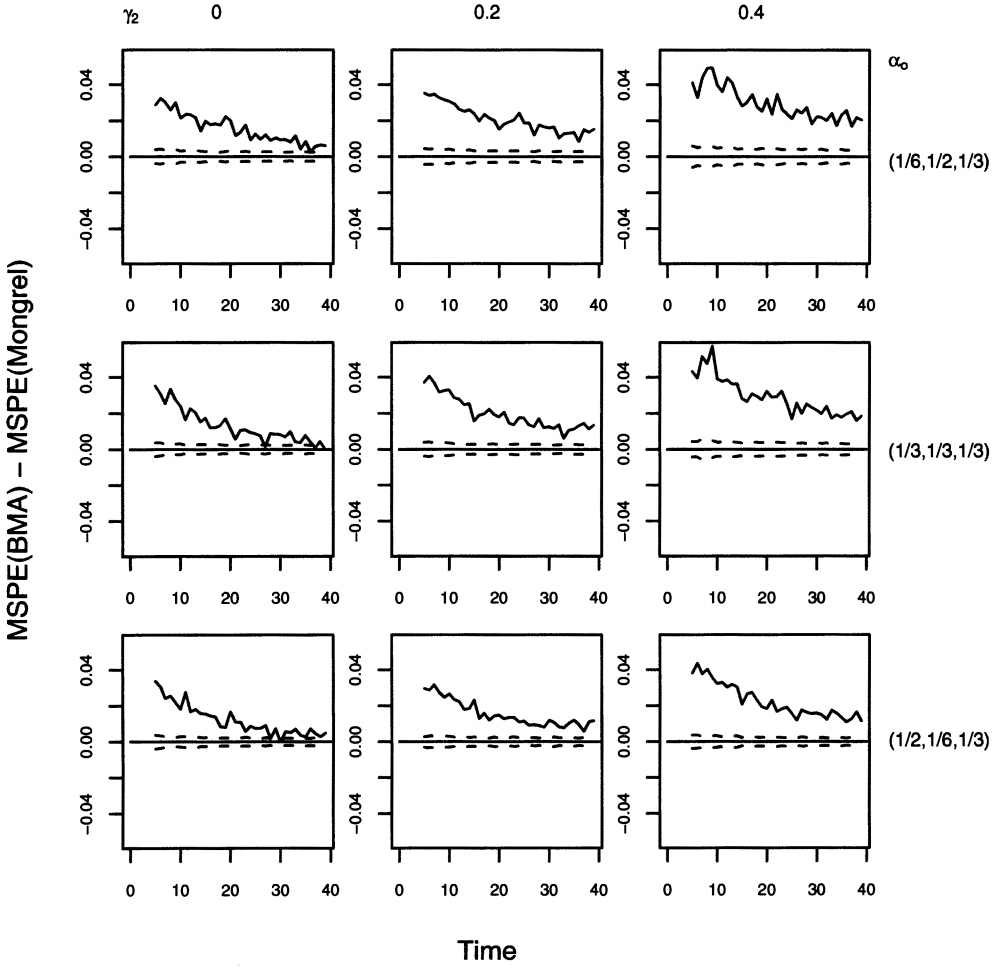

FIGURE 5: Performance of the mongrel procedure relative to BMA for different values of γ_2 and different prior model weights α_o when $\gamma_1 = 0.2$ and the model list has 3 models. The solid line is MSPE(BMA) − MSPE(Mongrel). The dashed lines represent ± 1 standard error for the difference.

5. DISCUSSION

We have proposed a new type of criteria, the mongrel risks, for selecting online predictors. The mongrel risk is novel in that it combines both model information and past empirical performance in evaluating candidate predictors. The application of the mongrel risk requires a rule for selecting the conditioning statistic $\mathbf{S}_n$. We have advocated an adaptive approach to selecting $\mathbf{S}_n$.

Our simulations show that an adaptive mongrel approach beats out BMA in small samples across a practically meaningful range of data-generators (values for γ_2) and prior model probabilities (values for α_0). Although our results are limited in scope, they provide compelling evidence that under a predictive criterion one can do better than always conditioning on all of the data, i.e. being Bayes. This is a question of quality versus quantity: When will a small number of highly informative statistics perform better than a larger number of less informative statistics?

An analytic treatment to assess the spectrum of scenarios over which the improvements are maintained is desirable, but hard to obtain, since the expressions needed to assess the performance of a mongrel procedure seem challenging. Moreover, because we are dealing strictly with small sample performance, we cannot appeal to asymptotic approximations.

Another setting in which the Bayes solution loses out to a competitor in a predictive setting is described by Clarke (2003). There, it is seen that as the approximation power of the model list is weakened, the performance of BMA relative to a cross-validation type of model averaging called *stacking*, deteriorates; i.e., as the data generator deviates from the elements of a model list, it is ever easier for stacking to beat BMA until both are so far wrong that model averaging is no help.

The fundamental components of an inference procedure are prior information, data, and the model list possibly equipped with a prior. The models provide the necessary framework for combining the information in the prior with the information in the data. To date, robustness of inference procedures has focused on sensitivity to atypical (i.e., bad) data primarily from a frequentist perspective or misspecification of priors from the Bayesian perspective. These two aspects of robustness are opposites: If the model list is good, i.e., approximates the data generator well, then a procedure that puts more weight on the prior than on the data is less sensitive to bad data but more sensitive to a bad prior. Since the mongrel procedure often discards some information, this makes it less sensitive to atypical data than the Bayes procedure. The gains seen in our results reflect the ability of the mongrel procedure to identify the atypical sequences for which the Bayes procedure has high risk and to select a better predictor for them. At the same time, the mongrel procedure seems to retain most of the usual robustness against choice of prior. This is consistent with the view that prior sensitivity is less important than the sensitivity to bad data.

The situation is less clear when a high level of model uncertainty is present and model sensitivity to the data may be the most difficult aspect to fix. Our results suggest that one can only beat Bayes when the model list is already pretty good given the data generator. This means that the model uncertainty is already unnaturally low. Indeed, suppose we define the diameter of a model list $\mathcal{M}$ to be

$$\mathcal{D}(\mathcal{M}) = \max_{i,j} d(i,j)$$

as i and j range over the models in $\mathcal{M}$ and d is a distance function. Then the contrast between the two three-model cases in Section 4 suggests that as $\mathcal{D}$ increases the degree by which the mongrel procedure beats out the Bayes procedure decreases until Bayes is better than the mongrel one. Note that the diameter is one way to express model uncertainty.

Since the distances between the data generator and the candidate models impact on how well averaging strategies, such as the mongrel, work, it is essential to throw out models that are sufficiently far wrong. One technique used in BMA is to throw out seemingly bad models based on the Bayes factor between a given model and the best model, i.e., the model achieving $\max_k m_k(\mathbf{y}_{(n)})$ for given $\mathbf{y}_{(n)}$, and then averaging over the remaining ones. This criterion is often combined with an Occam's window argument (Madigan & Raftery 1994) in which a simple model achieving a larger posterior probability than a more complex model discredits the complex model. In our limited computations, we found that applying Bayes factors to discard models degraded the performance of the mongrel procedure. Surprisingly, this usage also degraded the performance of BMA in our examples. We speculate that this occurs because Bayes factors intrinsically assess model fit, not predictive accuracy. The usual mean squared prediction error, conditional on all of the data, is an alternative to the Bayes factor that we are investigating as a

criterion for discarding bad models prior to applying the mongrel procedure.

APPENDIX: DERIVATION OF RELATIONS (6), (8), AND (9)

Derivation of (6)*:* The mongrel risk can be re-expressed as follows:

$$\begin{aligned}
\rho_i\left\{\widehat{Y}_{n+1}(\mathbf{S}_n)\right\} &= \mathrm{E}_{i;\mathbf{S}_n}\left\{Y_{n+1} - \sum_{k=1}^{K}\alpha_k(\mathbf{S}_n)\widehat{Y}_{k,n+1}\right\}^2 \\
&= \mathrm{E}_{i;\mathbf{S}_n}\left\{\sum_{k=1}^{K}\alpha_k(\mathbf{S}_n)\big(Y_{n+1} - \widehat{Y}_{k,n+1}\big)\right\}^2 \\
&= \mathbf{V}_{i;\mathbf{S}_n}\left\{\sum_{k=1}^{K}\alpha_k(\mathbf{S}_n)R_{k,n+1}\right\} + \left[\mathrm{E}_{i;\mathbf{S}_n}\left\{\sum_{k=1}^{K}\alpha_k(\mathbf{S}_n)R_{k,n+1}\right\}\right]^2 \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\alpha_j(\mathbf{S}_n)\alpha_k(\mathbf{S}_n)\mathbf{C}_{i;\mathbf{S}_n}\big(R_{j,n+1}, R_{k,n+1}\big) \\
&\quad + \left\{\sum_{k=1}^{K}\alpha_k(\mathbf{S}_n)\mathrm{E}_{i;\mathbf{S}_n}R_{k,n+1}\right\}^2.
\end{aligned}$$

Applying the identity $\mathbf{C}_{i;\mathbf{Z}}(X,Y) = \mathbf{C}_i(X,Y) - \mathbf{C}_i(X,\mathbf{Z})\{\mathbf{V}_i(\mathbf{Z})\}^{-1}\mathbf{C}_i(\mathbf{Z},Y)$, we obtain

$$\mathbf{C}_{i;\mathbf{S}_n}(R_{j,n+1}, R_{k,n+1}) = \mathbf{C}_i(R_{j,n+1}, R_{k,n+1}) - \Xi_{j,i}\Sigma_i^{-1}\Xi_{k,i}^{\top}, \tag{11}$$

where $\Xi_{k,i}$ is given by (7). Similarly, applying the identity $\mathrm{E}_{i;\mathbf{Z}}Y = \mathrm{E}_iY + \mathbf{C}_i(Y,\mathbf{Z})\{\mathbf{V}_i(\mathbf{Z})\}^{-1}(\mathbf{Z} - \mathrm{E}_i\mathbf{Z})$ gives

$$\mathrm{E}_{i;\mathbf{S}_n}R_{k,n+1} = \mathrm{E}_iR_{k,n+1} + \Xi_{k,i}\Sigma_i^{-1}(\mathbf{S}_n - \mu_i). \tag{12}$$

Substituting (11) and (12) into the above expression for $\rho_i\{\widehat{Y}_{n+1}(\mathbf{S}_n)\}$ gives (6).

Derivation of (8) *and* (9)*:* Under squared error loss and given data $\mathbf{Y}_{(n)}$ at time point n, the Bayes predictor for Y_{n+1} under model k with prior π_k is

$$\begin{aligned}
\widehat{Y}_{k,n+1} &= \mathrm{E}_{k;\mathbf{Y}_{(n)}}Y_{n+1} \\
&= \mathrm{E}_kY_{n+1} + \mathbf{C}_k\big(Y_{n+1}, \mathbf{Y}_{(n)}\big)\big\{\mathbf{V}_k\mathbf{Y}_{(n)}\big\}^{-1}\big(\mathbf{Y}_{(n)} - \mathrm{E}_k\mathbf{Y}_{(n)}\big).
\end{aligned}$$

The predictual arising from predicting Y_{n+1} using $\widehat{Y}_{k,n+1}$ is

$$\begin{aligned}
R_{k,n+1} &= Y_{n+1} - \big\{\mathbf{Z}_{k,n+1}^{\top}\mathbf{b}_k + \mathbf{Z}_{k,n+1}^{\top}\Gamma_k\mathbf{Z}_{k,(n)}^{\top}\Psi_{k,n}^{-1}\big(\mathbf{Y}_{(n)} - \mathbf{Z}_{k,(n)}\mathbf{b}_k\big)\big\} \\
&= \mathbf{u}_{k,n+1}^{T}\big(\mathbf{Y}_{(n+1)} - \mathbf{Z}_{k,(n+1)}\mathbf{b}_k\big)
\end{aligned} \tag{13}$$

with $\mathbf{u}_{k,n+1}^{T}$ as defined in (10). Applying expectation/covariance operators to $R_{k,n+1}$ in the form given by (13) yield (8) and (9).

ACKNOWLEDGEMENTS

The authors wish to thank Dr Rong Zhu for his assistance in coding the simulations. They are also grateful to the referees and to the Associate Editor for their detailed comments and suggestions. Partial support was provided by grants to the two authors from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

D. M. Allen (1974). The relationship between variable selection and prediction. *Technometrics*, 16, 125–127.

B. Clarke (2003). Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. *Journal of Machine Learning Research* 14, 683–712.

M. A. Clyde (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds). Oxford University Press, pp. 157–185.

A. P. Dawid (1984). Statistical theory: the prequential approach (with discussion). *Journal of the Royal Statistical Society Series A*, 147, 278–292.

A. P. Dawid (1992). Prequential data analysis. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (M. Ghosh & P. K. Pathak, eds.), IMS Lecture Notes – Monograph Series 17. Institute of Mathematical Statistics, Hayward, CA, pp. 113–126.

J. A. Hoeting, D. Madigan, A. E. Raftery & C. T. Volinsky (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, 14, 382–417.

D. Madigan & A. E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535–1546.

F. Mosteller & J. W. Tukey (1968). Data analysis, including statistics. In *Handbook of Social Psychology, Volume 2: Research Methods* (G. Lindzey & E. Aronson, eds.), Addison-Wesley, Reading, Massachusetts.

A. E. Raftery, D. Madigan, & J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179–191.

F. Seillier-Moiseiwitsch & A. P. Dawid (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, 88, 355–359.

K. Skouras & A. P. Dawid (1999). On efficient probability forecasting systems. *Biometrika*, 86, 765–784.

H. Wong (2000). *Small Sample Improvement Over Bayes Prediction Under Model Uncertainty.* Doctoral dissertation, Department of Statistics, The University of British Columbia, Vancouver, Canada.

Received 5 October 2001
Accepted 23 January 2004

Hubert WONG: hubert@hivnet.ubc.ca
Department of Healthcare and Epidemiology
University of British Columbia
Vancouver, British Columbia
Canada V6T 1Z3

Bertrand CLARKE: bertrand@stat.ubc.ca
Department of Statistics
The University of British Columbia
Vancouver, British Columbia
Canada V6T 1Z2