Bayesian Analysis (2013) **8**, Number 3, pp. 647–690

Prediction in $\mathcal{M}$-complete Problems with Limited Sample Size

Jennifer Lynn Clarke[*], Bertrand Clarke[†] and Chi-Wai Yu[‡]

Abstract. We define a new Bayesian predictor called the posterior weighted median (PWM) and compare its performance to several other predictors including the Bayes model average under squared error loss, the Barbieri-Berger median model predictor, the stacking predictor, and the model average predictor based on Akaike's information criterion. We argue that PWM generally gives better performance than other predictors over a range of $\mathcal{M}$-complete problems. This range is between the $\mathcal{M}$-closed-$\mathcal{M}$-complete boundary and the $\mathcal{M}$-complete-$\mathcal{M}$-open boundary. Indeed, as a problem gets closer to $\mathcal{M}$-open, it seems that $\mathcal{M}$-complete predictive methods begin to break down. Our comparisons rest on extensive simulations and real data examples.

As a separate issue, we introduce the concepts of the 'Bail out effect' and the 'Bail in effect'. These occur when a predictor gives not just poor results but defaults to the simplest model ('bails out') or to the most complex model ('bails in') on the model list. Either can occur in $\mathcal{M}$-complete problems when the complexity of the data generator is too high for the predictor scheme to accommodate.

Keywords: $\mathcal{M}$-complete, prediction, ensemble methods, basis selection, model selection, model list selection

1 Classes of Prediction Problems

Prediction problems naturally fall into three classes, namely $\mathcal{M}$-closed, $\mathcal{M}$-complete, and $\mathcal{M}$-open, based on the properties of the data generator (DG) (Bernardo and Smith 2000). Briefly, $\mathcal{M}$-closed problems are those where it is reasonable to assume that the true model is one of the models under consideration, i.e., the true model is actually on the model list (at least in the sense that error due to model mis-specification is negligible compared to any other source of error). This class of problems is comparatively simple and well studied.

By contrast, $\mathcal{M}$-complete problems are those where the DG has a true model that can be imagined but is not identifiable in any closed form. Inability to write a model explicitly may arise because the model is too complicated or because its constituent pieces are not known. The key point for an $\mathcal{M}$-complete problem is that it is plausible to assume that a true model – also called a 'belief model' – exists because this enables

[*] Division of Biostatistics, Department of Epidemiology and Public Health, University of Miami, Miami, FL, jclarke@biostat.med.miami.edu

[†] Departments of Medicine and of Epidemiology and Public Health, Center for Computational Science, University of Miami, Miami, FL, bclarke2@med.miami.edu

[‡] Department of Mathematics, Hong Kong University of Science and Technology maycw99@yahoo.com.hk

DOI:10.1214/13-BA826

its use in reasoning even if a prior cannot be meaningfully assigned in the usual way. For instance, if a true model exists a bias-variance decomposition can be developed, at least in principle, even when the true model is not explicitly known. Also, sometimes the true model in an $\mathcal{M}$-complete problem can be regarded as a limit point of models. That is, sometimes one can find approximations to the true model that provide serviceable inferences. If this approximation is good enough, the $\mathcal{M}$-complete problem becomes effectively $\mathcal{M}$-closed.

The $\mathcal{M}$-open class of problems is one step more elusive. $\mathcal{M}$-open problems are those in which the DG does not admit a true model. The DG is so complex (in some sense) that there is no true model that we can even imagine. For instance, one can regard the Collected Works of William Shakespeare as a sequence of letters. Unarguably this data set had a DG (William Shakespeare), but it makes no sense to model the mechanism by which the data was generated. One might try to use the first n letters to predict the $n+1$ letter and do better than merely guessing, but one should not expect such a predictor, or any model associated with it, to generate more great literature. The same point applies to the nucleotide sequence in a chromosome, the purchases of a consumer over time, and many other settings. In these cases, we are only able to compare different predictors without reference to a true model.

It is reasonable to write

$$\mathcal{M}\text{-open} \succ \mathcal{M}\text{-complete} \succ \mathcal{M}\text{-closed}, \tag{1}$$

where $\succ$ represents a decreasing complexity ordering. Bayes methods, in particular Bayes model averaging under squared error loss (L2-BMA, or just BMA when no confusion will result), are asymptotically optimal for $\mathcal{M}$-closed problems. This follows from the results in Section 7 of Skouras and Dawid (1998). More abstractly, this follows by observing that (i) the complete class theorem ensures any risk optimal solution is arbitrarily close to a Bayes solution for some prior, and (ii) observing that as sample size increases the effect of the prior drops out. An information-theoretic perspective on the optimality of L2-BMA is in Raftery and Zheng (2003) and Clarke et al. (2014).

At the other end of the complexity spectrum, no one really knows what techniques work well for $\mathcal{M}$-open problems. There is speculation that Frequentist techniques may have a role to play since the formulation of a prior is so problematic (Bernardo and Smith 2000). For instance, a model averaging technique called stacking finds model weights using a cross-validation criterion, see Wolpert (1992). Stacking often outperforms BMA in pseudo-$\mathcal{M}$-open settings (Clarke 2003). Separately, Clyde (2012) observes that releasing the usual stacking constraints (positive coefficients that sum to one) may be necessary to get predictively good model averages for $\mathcal{M}$-open problems. From a log-loss point of view, the Shtarkov solution (Shtarkov 1987) has also been extensively studied in the $\mathcal{M}$-open case, see Cesa-Bianchi and Lugosi (2006), but has not caught on partially because the conditions for it to exist are so narrow.

We focus on the intermediate case in (1), namely, $\mathcal{M}$-complete problems, for two reasons. First, this class of problems is the one most relevant to subject matter researchers. Most investigators believe there is a true model for the phenomenon they are studying even if they are only able to approximate it under narrow circumstances. Indeed, one can argue that the example used in Clyde (2012) is just a very complex $\mathcal{M}$-complete problem and not actually $\mathcal{M}$-open. Second, the $\mathcal{M}$-complete problem class seems to be relatively unexplored apart from the simplest cases where a really useful $\mathcal{M}$-closed approximation is available and possibly very complex cases as in Clyde (2012).

As suggested above, the class of $\mathcal{M}$-complete problems has a range of difficulty. Some $\mathcal{M}$-complete problems are just barely $\mathcal{M}$-complete in the sense that they can be extremely well approximated by models that can be written down conveniently. Essentially, this means the $\mathcal{M}$-complete problem is a limiting case of $\mathcal{M}$-closed problems and hence on the boundary between $\mathcal{M}$-closed and $\mathcal{M}$-complete. At the other extreme, some $\mathcal{M}$-complete problems are almost $\mathcal{M}$-open. The true model is so hard to approximate or the DG is so hard to simplify that the existence per se of a true model for the DG is no help. In this case one may attempt to reduce the mean squared error but the tradeoff between the variance and bias is trivial: Any effort to reduce bias increases the variance by a similar amount and any effort to reduce the variance increases the bias by a similar amount. The Doppler example below, see Section 3.3, may be an instance of this: Better modeling near zero costs as much in variance as it gains in reduced bias for moderate sample sizes. This may also be the case when the DG is a mixture of many subsidiary DGs (as in a mixture model) or when the factors influencing the DG are numerous, high-variance, latent, or complex. In these cases, the $\mathcal{M}$-complete problem is on the boundary between $\mathcal{M}$-complete and $\mathcal{M}$-open. One of the effects of the DG being too complex relative to the predictive scheme is that the predictors often become trivial. We call this the 'Bail-out effect', see Sections 3.3, 4.2 and 5.2. The 'Bail-in effect' – where predictors default to the most complex possibility – is also seen, see Section 4.2.

It is well known that the best predictors in a regression context are usually model averages. Recall that, under squared error loss, the L2-BMA given by the posterior mean of the predictors from the individual models on a model list is optimal, see Hoeting et al. (1999). Likewise, the Barbieri-Berger median model (BB) is predictively optimal in the sense that it provides an optimal approximation to the L2-BMA, see Barbieri and Berger (2004). BB can be regarded as a model average because it is constructed by including explanatory variables that have posterior probability at least one half when all models that contain them are considered. Stacking is another model average based on finding the weights for a model average by optimizing a cross-validation criterion (Wolpert 1992; Smyth and Wolpert 1999); its optimality in a squared error sense is given in Clyde (2012). Indeed, almost any model selection criterion can be converted into a model average by normalizing the values of the criterion for evaluating the models. This can be done with the Akaike information criterion yielding the AICMA (Johnson and Omland 2004). Other examples of model average predictors include Wong and Clarke (2004) for the $\mathcal{M}$-closed case and Shtarkov (1987) for the $\mathcal{M}$-open case, see Cesa-Bianchi and Lugosi (2006).

The main contribution of this paper is to propose a new predictor for $\mathcal{M}$-complete problems and verify through extensive computational comparisons that it compares favorably with four existing predictors given realistic sample sizes. The new predictor is called the posterior weighted median (PWM) and our evidence suggests that it can outperform L2-BMA, BB, stacking, and the AICMA, as well as a few other predictors, see Section 3.4. The degree of outperformance is slight in some cases and substantial in others, but typically holds for nested problems with orthogonal regressors until the difficulty of the $\mathcal{M}$-complete prediction problem approaches that of an $\mathcal{M}$-open problem. At this point it is unclear if any method for $\mathcal{M}$-complete problems is generally better than another.

All our examples use basis expansions (mostly orthogonal) in the explanatory variables to create predictors for a single random variable. We have chosen the specific examples presented here because we think they are representative of a large class of $\mathcal{M}$-complete problems where the models are naturally nested. One of the benefits of nesting is that we can easily interpret histograms of predictor selection when the predictors are specific models. We do this in Sections 4.2 and 5.1 for PWM and BB. Nesting can be achieved by using any shrinkage criterion for variable inclusion as a function of the decay parameter.

As a final introductory point, our comparison of the predictors described in Section 2 is empirical. Our focus is on how well a predictor given, say, n data points is able to predict the $n{+}1$ outcome; we call this the Final Predictive Error (FPE). The idea is that a correct measure of the success of a predictor must take into account all the sources of variability involved in the construction of the predictor and all of these sources are implicitly built into the comparison of the predicted value with the actual $n+1$ value over many uses of the predictor.

The rest of the paper is organized as follows. In Section 2 we present our predictors. In Section 3, we present computational results for these predictors in seven classes of univariate functions with three different model lists based on polynomial, Sine, and Fourier bases. We see that, in overall terms, PWM is the best method, closely followed by BB. We also see a first example of the ‘Bail-out Effect’. In Section 4 we look inside PWM and BB to examine which models each method chooses to make its predictions for $Y_{n+1}(\mathbf{x}_{n+1})$ and how the variance of the predictors depends on x_{n+1}. This is a way to ‘look inside the black box’ to understand what these predictors are doing, e.g., giving useful results, ‘bailing out’, or ‘bailing in’. In Section 5 we compare the five predictive methods on two multidimensional complex data sets. Finally, in Section 6, we discuss several important issues relating to the use of predictors including model list selection and complexity.

2 The Predictors

Assume we have data of the form $\mathcal{D}_n = (Y_1, \mathbf{X}_1)$, ..., $(Y_n, \mathbf{X}_n)$. Given $\mathbf{X}_{n+1} = \mathbf{x}_{n+1}$, a p-dimensional explanatory variable, our task is to predict $Y_{n+1}(\mathbf{x}_{n+1})$ by $\hat{Y} = \hat{Y}_{n+1}(\mathbf{x}_{n+1})$ assuming a signal-plus-noise model of the form $Y_{n+1}(\mathbf{X}_{n+1}) = f(\mathbf{X}_{n+1}) + \epsilon$, where ϵ is a noise term. Often we write $\mathbf{X}^{\text{new}}$ or $\mathbf{x}^{\text{new}}$in place of $\mathbf{X}_{n+1}$ or $\mathbf{x}_{n+1}$ to emphasize the fact that it is an input value that has not been seen before. We consider a class of distinct regression models $M_k = \{f_k(\mathbf{X} \mid \beta_k)\}$ for $k = 1, \ldots, K$ where k indicates the selection of explanatory variables from $\mathbf{X}$ used by model f_k and β_k is the parameter vector for the nonzero coefficients of model f_k. Denote the model list by $\mathcal{M} = \{M_1, \ldots, M_K\}$. Now, for linear regression models, we have $M_k : f_k(\mathbf{X} \,|\, \beta_k) = \mathbf{X}(f)^T \beta_k$ where the argument f indicates the selection of variables from $\mathbf{X}$ appearing in f_k.

In this notation, L2-BMA can be approximated as follows. First assign a weight to each model in $\mathcal{M}$ by using a Bayes information criterion (BIC) value. For model M_k, the weight is

$$w_k = \frac{\exp\{-0.5BIC_k\}}{\sum_{j\in\mathcal{M}} \exp\{-0.5BIC_j\}},$$

where BIC_j is the BIC value for model M_j. Following BB, the BIC is evaluated at the MLE, which bears a close relationship to the usual least squares estimator. Also, as in BB, the w_k's in (2) assume the prior over models is uniform. It is seen therefore that the w_k's do not depend on the within-model priors on the β_k's. For $\mathbf{x}^{\text{new}}$, we use

$$\hat{Y}_{\text{BMA}}(\mathbf{x}^{\text{new}}) = \sum_{k=1}^{K} w_k f_k(\mathbf{x}^{\text{new}} \mid \tilde{\beta}_k),$$

where (again following BB) $\tilde{\beta}_k$ is the posterior mean from $\mathcal{M}_k$ using the prior assigned to β_k.

By contrast, BB proposed a median based method for prediction by thresholding variable posterior inclusion probabilities and ensuring the result is a model within $\mathcal{M}$ so it can be used to give predictions. For $k = 1, \ldots, K$, if $f_k(\mathbf{X} \,|\, \beta_k) = \mathbf{X}(f)^T \beta_k$ the BB median model includes a given coordinate in $\mathbf{X}$ if and only if the set of models $\mathcal{M}_k$ including it has posterior probability at least one-half; the across-models posterior is found using (2). In the special case that the f_k's are nested, i.e., increasing with K, a prediction for Y_{n+1} at $\mathbf{x}^{\text{new}}$ is taken from the $f_{\hat{k}}(\mathbf{x}^{\text{new}}|\tilde{\beta}_{\hat{k}})$ satisfying

$$\hat{k} = \arg\max\{k : \sum_{j=k}^{K} w_j \geq .5, k = 1, \ldots, K\}.$$

More generally, for the definition of the median model to give valid predictions, the model list must have 'graphical model structure', i.e., any set of covariates that might end up in the median model must give a model in $\mathcal{M}$. We use $\hat{Y}_{\text{BB}}(\mathbf{x}^{\text{new}})$ to denote the predicted value of the response for $\mathbf{x}^{\text{new}}$ by BB.

To define a third predictor, note that, being a sum, BMA can be unduly affected by large or small values. Also, note that the BB median model makes predictions from the same model regardless of the value of $\mathbf{x}^{\text{new}}$. So, the new predictor we consider is an effort to reduce the variance of BMA (see the graphs in Section 4.3) by dropping terms that are too small or too large and to predict better than the BB median model by ensuring any predicted value is in the midrange of the model predictions. Our new predictor called the Posterior Weighted Median (PWM) will also have the benefit of being optimal in a conditional sense (see (2)). To define the PWM, put the predictions in increasing order and then choose one by writing

$$\hat{Y}_{\text{PWM}}(\mathbf{x}^{\text{new}}) = \underset{k=\{1,\ldots,K\}}{\text{med}}[w_k \diamondsuit f_k(\mathbf{x}^{\text{new}} \mid \tilde{\beta}_k)].$$

The notation $\diamondsuit$ means that $\underset{k=1,\ldots,K}{\text{med}}[w_k \diamondsuit f_k] = f_{(r)}$, where $r = \min\{j : \sum_{i \leq j,\, i=1,\ldots,K} w_{(i)} \geq 1/2\}$ and $w_{(i)}$ is the corresponding weight of the i^{th} smallest value of the f_k's, not w_k's, over $k = 1, \ldots, K$. As before, we follow BB in using the MLE's to define the BIC_k's in the w_k's but use the posterior means $\tilde{\beta}_k$ to make predictions. It is seen that PWM does not require 'graphical model structure'. An alternative (that we have not investigated) would be $\hat{Y}_{\text{FPWM}} = \underset{k=1,\ldots,K}{\text{med}}[f_k(\mathbf{x}^{\text{new}} \mid \tilde{\beta}_k)]$, a Frequentist form of PWM. However, FPWM should be slightly inferior to PWM as it does not use the information in the posterior weights.

PWM comes from optimizing an L^1 criterion using the across-model posterior. Specifically,

$$\begin{aligned} \hat{Y}_{\text{PWM}}(\mathbf{x}^{\text{new}}) &= \arg\min_u \sum_{k=1}^{K} |u - f_k(\mathbf{x}^{\text{new}}|\tilde{\beta})|w_k \\ &\approx \arg\min_u \sum_{k=1}^{K} |u - f_k(\mathbf{x}^{\text{new}}|\tilde{\beta})|W(M_k|\mathcal{D}_n) \end{aligned} \qquad (2)$$

where $W(\cdot|\mathcal{D}_n)$ is the (marginal) posterior distribution for the models in $\mathcal{M}$. Note that in (2) the optimum depends on $\mathbf{x}^{\text{new}}$. That is, different models may be used to give different predictions for different values of $\mathbf{x}^{\text{new}}$ – unlike BMA or BB. This heightened adaptivity may help PWM outperform other predictors just as the heightened adaptivity to the data improves the performance of the predictor in Wong and Clarke (2004).

By contrast, L2-BMA and BB both emerge from an L^2 optimality criterion. Hoeting et al. (1999) identifies the Bayes model average – literally an average of models in a distributional sense – as a posterior mean. Indeed, treating $\mathbf{x}^{\text{new}}$ as a parameter leads to the L2-BMA which is the familiar 'across-model posterior times predictor from the estimate of that model' form in (2). The posterior mean conditions on $\mathcal{D}_n$ giving a result independent of $\mathbf{x}^{\text{new}}$.

The BB predictor can be regarded as derived from L2-BMA. Specicially, Lemma 1 in Barbieri and Berger (2004) verifies that $\hat{Y}_{BB}(\mathbf{x}^{\text{new}})$ is the best single model L^2 approximation to the L2-BMA. This optimality conditions on the $y_1, \ldots, y_n$, involves an integration over the explanatory variables, and is independent of $\mathbf{x}^{\text{new}}$. The optimality also requires that the explanatory variables be orthogonal. Separately, Theorem 1 in Barbieri and Berger (2004) establishes the predictive optimality of the BB median model, i.e., BB is not just the best approximation to the BMA, it is also the best predictor in a squared error sense even though it is based on medians.

There is also an L1-BMA. Recall the BMA is an average of models which can be written as $p(Y_{n+1}|\mathcal{D}_n) = \sum_k W(k|\mathcal{D}_n)p(Y_{n+1}|k)$, see Hoeting et al. (1999). Just as L2-BMA is the posterior mean of $p(Y_{n+1}|\mathcal{D})$ and optimal in an L^2 sense, L1-BMA is the median of $p(Y_{n+1}|\mathcal{D})$ and optimal in an L^1 sense. However, while L2-BMA can be expressed in terms of the posterior means of the $p(Y_{n+1}|k)$'s, L1-BMA can not be readily expressed in terms of the posterior medians of the $p(Y_{n+1}|k)$'s. Moreover, it is seen that the L1-BMA depends on all terms in the BMA, making it different from BB and PWM which give predictions at each $\mathbf{x}^{\text{new}}$ from one model only. We did not investigate the performance of L1-BMA because we expected it to lose efficiency quickly as the model list increased (even L2-BMA loses efficiency in a relative sense; see Section 4.3).

It is easy to see that the more the across-model posterior concentrates on a single model the closer the predictions from PWM, BB, and PWM will be. In fact, PWM and BB will coincide when there is a single model that has posterior probability at least one-half. In such a case the model will dominate the terms in the L2-BMA and dominate the distribution of the L1-BMA.

For the sake of completeness, we also examine the performance of two Frequentist model averaging strategies. The first is a variation on L2-BMA formed by using the Akaike Information Criterion (AIC) to form a model average (AICMA). Recall that the BIC uses a $c \log n$ penalty where $2c$ is the number of parameters. For AICMA, we merely replace the BIC weights in (2) with AIC weights (still using the MLE's $\hat{\beta}_k$). Thus, for model M_k, we set

$$w'_k = \frac{\exp\{-0.5AIC_k\}}{\sum_{j \in \mathcal{M}} \exp\{-0.5AIC_j\}}, \tag{3}$$

where AIC_j is the AIC value for the model M_j. For a new observation $\mathbf{x}^{\text{new}}$, we use

$$\hat{Y}_{\text{AICMA}}(\mathbf{x}^{\text{new}}) = \sum_{k=1}^{K} w'_k f_k(\mathbf{x}^{\text{new}} \mid \tilde{\beta}_k). \tag{4}$$

The motivation for AIC in model selection is predictive, however, it is well known that for $\mathcal{M}$-closed model selection AIC is often unsuccessful, see Kashyap (1980), Shibata (1983). However, Leung and Barron (2006) demonstrates desirable theoretical properties of the AIC (and AIC-like model selection procedures) for model averaging and the usefulness of AIC for $\mathcal{M}$-complete problems is noted in Shibata (1981) and Clarke et al. (2014). Moreover, AICMA has been used to good effect in $\mathcal{M}$-complete problems in environmetrics (Johnson and Omland 2004; Symonds and Moussalli 2010).

The second Frequentist technique we include in our comparisons is stacking, a model averaging procedure motivated by cross-validation. Let $\hat{f}_k^{-j}(x)$ denote the prediction from model k evaluated at x where the coefficients in the f_k are estimated using MLE's based on all the data outside a holdout set, say D_j. Using five-fold cross-validation, i.e., $j = 1, \ldots, 5$, the stacking weights $\lambda_{k,opt}$ for the f_k's are obtained by a quadratic minimization

$$\min_{\lambda_1,\ldots,\lambda_K} \sum_{j=1}^{5} \sum_{i \in D_j} (y_i - \sum_{k=1,\ldots,K} \lambda_k \hat{f}_k^{-i}(x_i))^2,$$

where the λ_k's are assumed positive and sum to one. For a new observation $\mathbf{x}^{\text{new}}$

$$\hat{Y}_{\text{STK}}(\mathbf{x}^{\text{new}}) = \sum_{k=1}^{K} \lambda_{k,opt} f_k(\mathbf{x}^{\text{new}} \mid \tilde{\beta}_k). \tag{5}$$

Stacking was proposed in Wolpert (1992) and studied in Breiman (1996) and tends to do well in problems where the DG is not near a model that can be readily conceptualized from within the model list (Clarke 2003). It is likely that stacking will be optimal whenever cross-validation is optimal. For instance, Shao (1997) shows an optimality property of leave-one-out cross validation in the $\mathcal{M}$-complete ('Class I') case.

3 Comparing the Methods

To understand the behavior of PWM, BB, AICMA, STK, and BMA for small to moderate sample sizes we look at how they perform as a function of their inputs for randomly generated target functions from well-defined classes with a variety of loss functions. We do this in the context of a large scale computational comparison of their predictive performance. The basic structure of the simulations is the following. Fix a function class $\mathcal{F}$ on an interval $[a, b]$. Draw a function f from $\mathcal{F}$ at random and then draw $n+1$ values x_1,...,x_{n+1} of the single explanatory variable X from $[a, b]$. Find $f(x_1)$,...,$f(x_{n+1})$ and then let $y_i = f(x_i) + \epsilon_i$ for $i = 1, \ldots, n+1$ where the ϵ_i's are IID outcomes of a noise term $\epsilon \sim N(0, \sigma^2)$. We take $\sigma = 1$ unless specified otherwise. Then, using the first n data points we form each of the five predictors, generate a prediction $\hat{y}_{n+1}$ for y_{+1} at x_{n+1}, and look at the differences $(\hat{y}_{n+1} - y_{n+1})$ over repeated iterations. Apart from the random selection from $\mathcal{F}$, this is a standard simulation setting for regression problems.

In the next subsections we describe our seven function classes and the specifics of the computational setting, and then proceed to present our results. We conclude this section with a general interpretation of the results.

3.1 Nested Function Classes

To define the predictors, consider an ensemble of terms $\mathcal{E} = \{e_1, \ldots, e_K\}$ to form K nested linear regression models of the form

$$Y = f(X) + \epsilon = \sum_{u=1}^{k} \beta_u e_u(X) + \epsilon$$

where $k = 1, \ldots, K$. Here, $\mathcal{E}$ will be the first K elements of one of three bases: Chebyshev polynomials, the sine basis, and the Fourier basis consisting of both sines and cosines. That is, in all cases, we use a nested model list consisting of the first k elements of these three orthogonal bases in their usual order; when k is odd, the element chosen from the Fourier basis is $\sin kx$ and when $(k+1)$ is odd the element chosen from the Fourier basis is $\cos kx$. Since we omit the constant term, $K = 29$ in all of our simulations as was used in Barbieri and Berger (2004).

For all five predictors and in all models the β_u parameters are estimated by posterior means under IID $N(0, u^a)$ priors where $a = \pm 1, \pm 3$; the negative values correspond to sparsity while the positive values correspond to diffuseness. Thus, the within-model priors affect the non-Bayesian model averages via the $\tilde{\beta}_k$'s; coupled with using MLE's to find the model weights in all five cases, this makes the predictors more comparable. For the Bayesian predictors, we use a simple uniform prior across the models, invoking the Principle of Insufficient Reason (Bernardo and Smith 2000). With these specifications we can use the data (x_i, y_i) $i = 1, \ldots, n$ and x_{n+1} to obtain $\hat{Y}_{\text{PWM}}(\cdot)$, $\hat{Y}_{\text{BB}}(\cdot)$, $\hat{Y}_{\text{BMA}}(\cdot)$, $\hat{Y}_{\text{STK}}(\cdot)$, and $\hat{Y}_{\text{AICMA}}(\cdot)$ for Y_{n+1}.

We compare the performance of PWM, BB, BMA, AICMA, and STK on seven function classes. They are as follows.

- *Trigonometric (Trig):* This class of functions is a collection of Fourier expansions of varying length and coefficients. It is defined to be of the form

$$f_{\mathbf{a},\mathbf{b},d}(x) = a_0 + \sum_{j=1}^{d} a_j \sin(jx) + b_j \cos(jx) \quad x \in [0,1],$$

  $d \sim \text{DUnif}[1,20]$, $a_j \sim \text{Unif}[0,1]$ for $j = 0, \ldots, d$, and $b_j \sim \text{Unif}[0,1]$ for $j = 1, \ldots, d$. (Here, $\text{DUnif}[a,b]$ means the discrete uniform on the integers $a, a+1, \ldots, b-1, b$.)

- *Neural Net (NN):* This class of functions consists of single hidden layer neural networks with varying numbers of nodes and varying coefficients at the nodes. It is defined as follows:

$$f_{\mathbf{a},\mathbf{b},\mathbf{c},r} = b_0 + \sum_{j=1}^{r} \frac{a_j}{1 + e^{b_j + c_j x}} \quad x \in [0,1],$$

  where $a_j, b_j, c_j, b_0 \sim \text{Unif}[-10, 10]$ for $j = 1, \ldots, r$, and $r \sim \text{DUnif}[2, 10]$.

- *Bumps:* This class of functions puts a random number of bumps at random locations with random heights. It is defined as follows, cf. Donoho and Johnstone (1994). Let $d \sim \text{DUnif}[7, 15]$:

$$f_{\mathbf{x},\mathbf{h},\mathbf{w},d}(x) = \sum_{j=1}^{d} \frac{h_j}{(1 + |x - \ell_j|/w_j)^4} \quad x \in [0,1]$$

  $\ell_j \sim \text{Unif}[0,1]$, $w_j \sim \text{Unif}[5/1000, 30/1000]$ and $h_j \sim \text{Unif}[2,6]$, for $j = 1, \ldots, d$.

- *Blocks:* This class of functions is a collection of step functions with varying heights and bin widths. It is defined as follows, cf. Donoho and Johnstone (1994). Let $d \sim \text{DUnif}[2, 10]$ and

$$f_{\mathbf{h},\mathbf{b},d}(x) = \sum_{j=1}^{d} h_j (1 + \text{sign}(x - b_j))/2 \quad x \in [0,1],$$

  with $h_j \sim \text{Unif}[-5, 5]$ for $j = 1, \ldots, d$, and $b_j \sim \text{Unif}[0,1]$ for $j = 1, \ldots, d$.

- *Treed:* This class of functions is a generalization of the *Block* class. Treed consists of one dimensional trees with varying leaves having polynomials of varying degrees:

$$f_{\ell,\mathbf{a},\mathbf{c_0},\mathbf{c_1},\mathbf{c_2},d}(x) = \sum_{j=1}^{d+1} \chi_{a_{(j-1)},a_{(j)}}(x) \sum_{u=0}^{\ell_j} c_{j,u} x^u \quad x \in [0,1],$$

  where $d \sim$ DUnif$[2,10]$, $\ell_j \sim$ DUnif$[0,2]$ for $j = 1,\ldots,d$, $a_j \sim$ Unif$[0,1]$ for $j = 1,\ldots,d$ (with $a_0 = 0$ and $a_{d+1} = 1$), and all $c_{j,u} \sim$ Unif$[-5,5]$. The $a_{(j)}$'s are order statistics and $\chi_{a_{(j-1)},a_{(j)}}(\cdot)$ is the indicator function for the region between the percentiles $a_{(j-1)}$ and $a_{(j)}$

- *Logarithm (Log):* This class of functions is defined to be of the form

$$f_{a,b}(x) = -a\log(b(1-x)) \quad x \in (-1,1),$$

  where $a \sim$ Unif$[-5,5]$ and $b \sim$ Unif$[1,5]$, cf. Barbieri and Berger (2004).

- *Doppler:* This class of functions is of the form

$$f_{a,b}(x) = \sqrt{x(1-x)}\sin\frac{b\pi(1+a)}{x+a} \quad x \in [0,1]$$

  where $a \sim$ Unif$[0,.5]$ and $b \sim$ Unif$[-10,10]$, cf. Donoho and Johnstone (1994).

The first three classes are smooth, with the first two having 'nice' appearances even though NN's are typically more complex than our Trigonometric functions. Bumps functions are also 'nice' because they are well-behaved almost everywhere; the only difficult points are the fourth order cusps at random locations. Representatives of these function classes are shown in Figure 1.

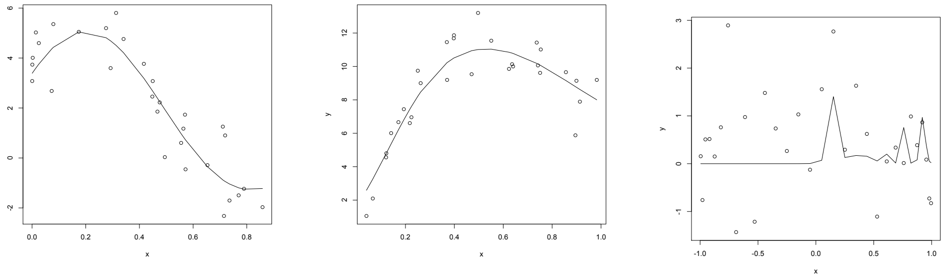

Figure 1: Typical representatives of the Trig, NN, and Bumps function classes each on their own domain.

The last four classes are considered not 'nice'. The Blocks and Treed classes are not smooth because of jump discontinuities; in fact Treed functions are a generalization of Block functions. The Log and Doppler function classes have regions (not just isolated points) where they are difficult. Log has a region of extremely rapid increase and an asymptote; Doppler has a region of rapid oscillations that converge to the origin. Representative elements of these classes are plotted with data in Figure 2.

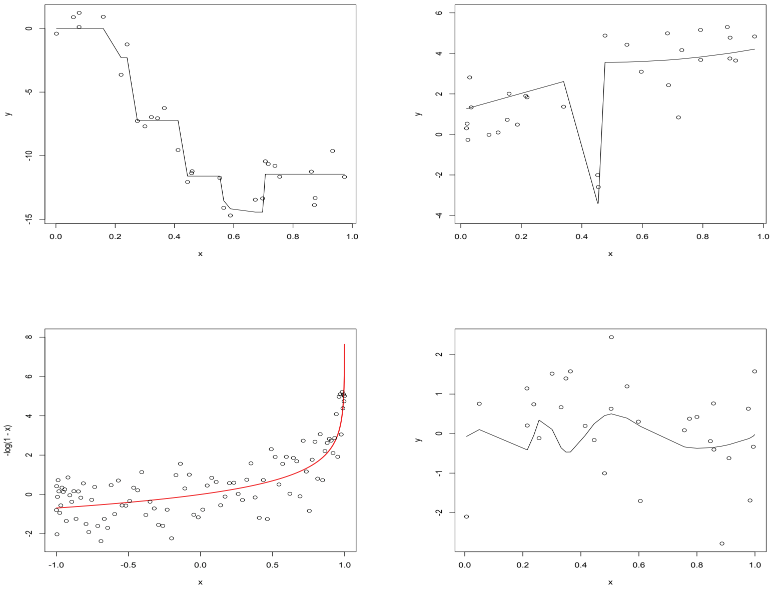

Figure 2: Typical representatives of the Block and Treed (top) and Log and Doppler (bottom) function classes each on their own domain.

3.2 Computational Setting

Using basis representations to predict outcomes from DG's based on these seven classes mimics $\mathcal{M}$-complete problems. Aside from a few cases in which the Trig class is used with the Sine or Fourier basis, the true DG is not accessible except in the limit of increasing size of model lists, here presumed to be of the same form e.g., including more basis elements. Indeed, for the not 'nice' functions there are entire regions on which an approximation from the ensemble will likely be poor regardless of ensemble size. This mimics believing the DG is not on the list and cannot be reliably approximated everywhere, and loosely corresponds to having only the conceptual existence of a model on a region.

In all our comparisons, the performance criterion is the final predictive error (FPE)

$$FPE = \sum_{u=1}^{R} |y_u(x_{u,n+1}) - \hat{y}_u(x_{u,n+1}|(y_{u,1}, ..y_{u,n}, x_{u,1}, ..., x_{u,n}))|^p \qquad (6)$$

where R is the number of iterations, $n+1$ is the sample size, and p defines the distance. We select a function class and draw R random functions from this class. For the u-th random function we randomly draw n data points $(x_{u,i}, y_{u,i})$ for $i = 1, \ldots, n$ (with $N(0,1)$ error). The first n data points and the prior are used to form the predictor $\hat{y}_u(\cdot)$, which we use to predict the $(n+1)$-th value of the u-th function at $x_{u,n+1}$.

Clearly, (6) is a comprehensive measure of the predictive error for the last observation because it is averaged over all the inputs to the predictive process (the DG, the random 'design points' $x_1, \ldots, x_n$, and the error term). The FPE also depends on the priors on the parameters in the models since the parameter estimates are required for all the predictors. The Bayesian predictors also depend on the prior on the model list, but as this is always uniform there is no basis for assessing its effect.

In this Section, we limit our attention to the case $p = 1$, $n = 100$, $a = -1$ and $R = 2000$. First, to choose $p = 1$, we started by only looking at $p = 1, 1.5, 2$ on the grounds that the optimality properties of the predictors only occurred in L^1 or L^2 and $L^{1.5}$ was the midpoint. Indeed, PWM is optimal in an L^1 sense while BB, BMA, STK, and AICMA have their optimality properties in L^2. (AICMA actually has its optimality property in terms of the relative entropy which is locally L^2 and has other L^2-like properties Clarke et al. (2014).) Then we observed that over $p = 1, 1.5, 2$ the performance of the methods as measured by the rankings of the FPE's was constant for our simulations. That is, if PWM did better than BB which did better than BMA for $p = 1$, the same ranking was observed for $p = 1.5$ and $p = 2$. So, looking at any one value of $p = 1, 1.5, 2$ was enough and $p = 1$ is a reasonable choice. Had we permitted $p >> 2$, the rankings of the methods in terms of FPE might not have been constant since, as p grows, more weight is put on large (but infrequent) deviations. Later, in Section 5.2, the choice of loss function will be informative as data accumulate with the real data set Fires.

Second, to choose $n = 100$, note that the comparative performance of methods matters most in small to moderate sample sizes. However, we wanted n large enough that differences in the methods could be plausibly attributed to the methods rather than random fluctuations. We did not consider large n because the Bayesian methods (BMA, BB, PWM) are asymptotically predictively optimal in an FPE sense and there are conditions under which AIC and probably STK (since it's based on cross-validation) are asymptotically consistent and hence may also give asymptotically optimal predictors. In our testing with $n = 30, 50, 100$ we found that the relative performance of the methods was mostly unaffected by n, but less stable when $n = 30$. Admittedly, there were a few cases, mostly with small n, in which the priors on the parameters seemed to have a small effect on the relative performance of the methods. Also, there were a few cases, with both $n = 30$ and $n = 100$, mostly with the hardest function classes, in which one or another of the non-PWM methods did best, usually only by a very small amount. We regarded this either as random variation or as representing the approach of the $\mathcal{M}$-complete problem to an $\mathcal{M}$-open problem, a different phenomenon. Thus, we set $n = 100$ as it was the largest plausible pre-asymptotic value that gave clear comparisons among the methods.

Third, at least for $n = 100$, it was enough to set $a = -1$, i.e., impose mild pressure towards sparsity, for the purposes of presenting representative tables of the output. We examined $a = \pm 1, \pm 3$. The choices of $a = -1, -3$ were motivated by Barbieri and Berger (2004) who chose negative values of a to impose sparsity. As a decreases, the prior makes large values (positive or negative) of coefficients on higher order terms less likely. This sparsity often gives better performance in practice when the target function is well approximated by a finite series expansion. On the other hand, a larger value of

a such as $a = 1, 3$ is a better representation of lack of knowledge about the coefficients. The cost is that positive values of a make large coefficients on high order terms more likely and hence there will be more instability associated with higher order terms, a property that might be undesirable. Since we are only looking operationally at the predictors, we ignore this issue apart from noting that the value of a matters more with small n than large n. So, while PWM performs better than the other four methods (in the sense of (6)) for $n = 100$, there are cases with smaller n where BB or BMA seems slightly better than PWM, e.g., $a = 3$ and $n = 30$ with Log or Doppler and the Sine basis. However, by the time $n = 100$ the effect of the priors on the relative performance of the methods is negligible for the first six of the seven function classes; the behavior of the predictors with the seventh class is more complex.

Fourth, we chose $R = 2000$ because this was essentially the largest number of iterates that we could compute across cases when $n = 100$ in a relatively short time frame. In fact, in the tables below for $a = -1$, the FPE's are an average of five FPE's over runs of 2000 iterates, hence 10,000 iterations total. The standard error (SE) reported is the SD of the five average FPE's, each with $R = 2000$, divided by $\sqrt{5}$. We report SE's for the cases with $a = -1$ because the $a = -1$, $n = 100$ cases were generally representative of the typical behavior of the predictors.

In the next subsection we present our simulation results. For six of the function classes (Trig, NN, Bumps, Block, Treed, and Log) we argue that PWM outperforms BB slightly and both PWM and BB outperform BMA, AICMA and STK (apart from two exceptions with the Chebyshev basis). There are two steps to this. The first step is to see that both PWM and BB outperform BMA, AICMA, and STK because in most cases the differences between FPE's for different predictor settings are large enough to be detected by arguments based on confidence intervals. (In the present context, confidence interval arguments are valid because simulation results are based on the repeated sampling definition of probability.) For instance, we can form $\pm 3SE$ intervals around mean FPE's (over five runs of 2000) of PWM or BB and see they are often separated from the corresponding $\pm 3SE$ intervals around mean FPE's of BMA, AICMA, and STK. For instance, we can compare PWM to AICMA for the Sine basis by observing

$$\begin{aligned}&\left[\overline{FPE(PWM)} \pm 3SE(FPE(PWM))\right]\\ \cap &\left[\overline{FPE(AICMA)} \pm 3SE(FPE(AICMA))\right] = \phi,\end{aligned} \tag{7}$$

see Table 2. Expression (7) means that the true FPE for PWM is almost certainly smaller than the true FPE for AICMA. This argument holds for all but a very few of the comparisons of PWM or BB to BMA, AICMA, or STK as can be seen in the first six tables in Section 3.3.

The second step is to see that PWM is slightly better than BB. These are generally cases where the difference between mean FPE's is too small to be revealed by (7). These cases result, in our view, because we did not have the computing power to do enough simulations to get small enough SE's for use in (7). For these cases, we use a Bayes testing argument. For example, let us compare PWM to BB even when the SE's are not small enough to reveal a difference as in the case with the Sine basis, the Trig functions, $a = -1$, $n = 100$, and $p = 1$; see Table 2. Consider testing

$$\begin{aligned} &\mathcal{H}_0 : E(FPE(PWM)) \geq E(FPE(BB)) \\ vs. \quad &\mathcal{H}_1 : E(FPE(PWM)) < E(FPE(BB)), \end{aligned} \tag{8}$$

using the simulation data in Table 1 where in each run of 2000 both methods used the same seed in the random number generator. Instead of (7), a de facto t-test, we can use a Bayesian sign test by writing $Z_i = 1$ for $FPE(BB) - FPE(PWM) > 0$ and $Z_i = 0$ otherwise, for $i = 1, \ldots, 5$. Regarding $Z = \sum_{i=1}^{5} Z_i \sim \text{Bin}(5, \pi)$ and using a $\text{Beta}(1, 1)$ (uniform) prior on π, we find that the posterior probability is $W(\pi > .5|Z = 5) = .89$ and the Bayes factor is BF $= 8.1$, meaning that we have good evidence to reject the null, i.e., we are led to conclude that PWM has a lower mean FPE than BB does. In the case that only four of the Z_i's are one, $W(\pi > .5|Z = 4) = .74$ and the Bayes factor is BF $= 2.84$. Bearing in mind that $Z = 4, 5$ holds for almost all the comparisons of PWM to BB for the first six function classes and further runs would only reinforce this pattern, it is safe to conclude quite generally that PWM has a lower FPE than BB even when the SE's are not small enough to show it. In fact, the patterns observed are strong enough that even if one took the multiple comparisons issues into account, the conclusions would be unchanged.

Run	FPE(PWM)	FPE(BB)	Z
1	1.834	1.835	1
2	1.840	1.841	1
3	1.789	1.795	1
4	1.794	1.795	1
5	1.806	1.813	1

Table 1: Results from five runs of 2000 iterations using the Sine basis, Trig class, $a = -1$, $n = 100$, and $p = 1$. These runs gave the (average) FPE's and standard errors for the Sine basis for PWM and BB in Table 2.

We comment that for the seventh function class, Doppler, the results are inconclusive. We attribute this to Doppler being closer to the $\mathcal{M}$-complete–$\mathcal{M}$-open boundary than are the other function classes. In Section 3.4 we discuss other methods and try to summarize the overall findings from these simulations.

3.3 Presentation of Simulation Results

We consider seven function classes of which there are three 'nice' smooth classes (as above); the remaining not 'nice' classes we divide into two classes with jump discontinuities, and two 'hard' function classes. In the first five classes – 'nice' and jump discontinuities – PWM generally does best. It also usually does best with the sixth (Log) class. However, for the seventh (Doppler) class there is no consistent pattern.

'Nice', Smooth Functions

Here we present our computational comparisons for the Trig, NN, and Bumps function classes. For the sake of concision, for all five predictors we only show tables for the three bases for $n = 100$, $a = -1$, $p = 1$, and the mean and SE of the FPE's from five runs of 2000 replications ($SE(FPE) = SD(FPE)/\sqrt{5}$ where $SD(FPE)$ is the SD of the five FPE values). As a practical matter, the very few runs in which there was a (very small) prior effect can be ignored as negligible in contrast to random variation. In all cases, the relative ordering of the five methods in terms of FPE's is representative of the other values of a, p and n.

Table 2 shows the FPE's and their SE's for the Chebyshev, Sine and Fourier basis for PWM, BB, L2-BMA, AICMA, and STK using the Trig function class. Bold font indicates the best performance in each row. It is seen that for the Chebyshev basis STK is the clear winner, while for Sine and Fourier PWM and BB are the clear winners even though their SE's are too large to separate them. The appearance of STK with Chebyshev as the overall winner in terms of low FPE's, with PWM and Fourier second, may seem surprising until one realizes that the Sine and Fourier bases are far more capable of representing elements of the Trig function class so we expect more bias when the Chebyshev basis is used. Hence, we expect STK to perform well since it is known to work better than some methods in the presence of bias Clarke (2003). When the bias is smaller, as it would be with Sine and Fourier, STK performs poorly as expected. Note that when the basis has low bias, it is seen that PWM and BB are better than BMA, even if not as good in terms of FPE as STK with Chebyshev. As the figures in Section 4.3 show, L2-BMA performs poorly because it is less stable than PWM or BB. This is typical when comparing mean-based to median-based methods. AICMA performs similarly to L2-BMA in many cases, possibly because both are essentially the result of L^2 optimizations.

Trig. functions	Bayes			Frequentist	
Method	PWM	BB	BMA	AICMA	STK
FPE Cheb, a=-1, n=100	1.01	1.013	2.697	2.682	**.825**
SE(FPE)	.006	.005	.030	.022	.008
FPE Sin, a=-1, n=100	**1.813**	1.816	2.717	2.767	14.457
SE(FPE)	.010	.010	.025	.025	.130
FPE Four, a=-1, n=100	**1.228**	1.229	2.683	2.686	14.53
SE(FPE)	.012	.012	.030	.05	.143

Table 2: FPE's and SE's for five methods, three bases, $a = -1$, $n = 100$, and $p = 1$ for the Trig function class.

Table 3 shows the results for the NN function class. For each basis PWM is numerically best, although BB is within one SE of PWM in all these cases. BMA, AICMA, and STK have credible or confidence sets that are disjoint (or nearly so) from credible sets for PWM and BB. The overall best method is PWM with Fourier and the worst basis is clearly Sine, perhaps because it has relatively more rapid oscillations and therefore worse bias.

NN's	Bayes			Frequentist	
Method	PWM	BB	BMA	AICMA	STK
FPE Cheb, a=-1, n=100	**.954**	.957	8.676	8.559	2.871
SE(FPE)	.007	.007	.068	.056	.024
FPE Sin, a=-1, n=100	**8.526**	8.528	8.655	8.814	8.769
SE(FPE)	.063	.063	.063	.063	.063
FPE Four, a=-1, n=100	**.939**	.941	8.631	8.579	8.754
SE(FPE)	.005	.005	.072	.063	.072

Table 3: FPE's and SE's for five methods, three bases, $a = -1$, $n = 100$, and $p = 1$ for the NN function class.

Table 4 shows the FPE's and their SE's for the Chebyshev, Sine and Fourier basis for PWM, BB, L2-BMA, AICMA, and STK using the Bumps function class. The results are qualitatively similar to those for the NN class. However, (1) the Sine basis does not perform as badly relative to the other two bases, (2) BB ties PWM for lowest FPE with Sine, and (3) few of the FPE's are distinguishable by using the SE's. The smallest FPE occurs for PWM and Fourier at 1.034 and the largest occurs for AICMA and Sin at 1.10; these two are distinguishable by using the SE's. The relatively small differences between most approaches implies that none of the methods really do well – perhaps because they try to capture the behavior of the functions away from the random bumps.

Bumps	Bayes			Frequentist	
Method	PWM	BB	BMA	AICMA	STK
FPE Cheb, a=-1, n=100	**1.042**	1.044	1.078	1.064	1.090
SE(FPE)	.009	.008	.013	.011	.008
FPE Sin, a=-1, n=100	**1.050**	**1.050**	1.078	1.1	1.084
SE(FPE)	.008	.008	.011	.008	.006
FPE Four, a=-1, n=100	**1.034**	1.035	1.062	1.07	1.090
SE(FPE)	.009	.009	.013	.008	.008

Table 4: FPE's and SE's for five methods, three bases, $a = -1$, $n = 100$, and $p = 1$ for the Bumps function class.

Jump Discontinuities

Here we present our computational comparisons for the Blocks and Treed function classes. The tables we present are entirely analogous to those of the previous section.

In Table 5 it is seen that for each basis PWM does best, although its FPE cannot be separated from the FPE's for BB. However, L2-BMA, AICMA, and STK can be separated by using confidence or credible sets based on the SE's and perform worse than PWM or BB. Moreover, the Sine basis is also seen to be the worst basis possibly for the

Block	Bayes			Frequentist	
Method	PWM	BB	BMA	AICMA	STK
FPE Cheb, a=-1, n=100	**1.146**	1.149	3.610	3.610	3.809
SE(FPE)	.008	.008	.034	.034	.015
FPE Sin, a=-1, n=100	**2.295**	2.306	3.686	3.61	3.768
SE(FPE)	.011	.012	.015	.025	.026
FPE Four, a=-1, n=100	**1.256**	1.261	3.656	3.629	3.768
SE(FPE)	.007	.007	.034	.020	.026

Table 5: FPE's and SE's for five methods, three bases, $a = -1$, $n = 100$, and $p = 1$ for the Blocks function class.

same reason as given in Table 3. The Fourier basis is less oscillatory so its performance is intermediate between Chebyshev and Sine. The Chebyshev basis may do better with discontinuities because it has no regular oscillations and hence can better locate jumps through shifts. The lowest numerical FPE occurs for PWM with Chebyshev. The results in Table 6 for the Treed class are similar to those from Table 5.

Treed	Bayes			Frequentist	
Method	PWM	BB	BMA	AICMA	STK
FPE Cheb, a=-1, n=100	**1.015**	1.018	3.809	2.735	2.873
SE(FPE)	.022	.022	.055	.017	.025
FPE Sin, a=-1, n=100	**2.536**	**2.536**	2.802	2.807	2.863
SE(FPE)	.008	.009	.010	.017	.022
FPE Four, a=-1, n=100	**1.163**	1.167	2.743	2.768	2.871
SE(FPE)	.014	.014	.011	.008	.024

Table 6: FPE's and SE's for five methods, three bases, $a = -1$, $n = 100$, and $p = 1$ for the Treed function class.

Two Hard Function Classes

Here we present our computational comparisons for the Log and Doppler function classes. The tables we present are entirely analogous to those of the previous section.

Table 7 shows that for $n = 100$, $a = -1$, Log, and the Chebyshev basis AICMA gives the best results. AICMA puts a relatively small penalty on including more terms enabling it to make predictions from a slightly larger and hence more accurate model (in the case of mild sparsity, i.e., $a = -1$). In contrast to the Trig function class, where STK did best with Chebyshev, it may be that the inclusion of more terms by AICMA permits more weight on higher order terms. The resulting model could mimic the behavior of Log near its asymptote as considering the Taylor expansion of Log would suggest. However, for $a = 1, 3$, Log, $n = 100$, and Chebyshev, PWM does best and slightly better than BB. (The difference persists over five runs.) In the case $a = -3$, BB does best and slightly better than PWM. For $a \neq -1$, the other three methods, AICMA, STK, and BMA, the FPE's differ more than could be explained by the SE's.

For Log and the Sine basis with $a = \pm 3, \pm 1$, BB and PWM are nearly equivalent; see Table 7. However, in our runs PWM tended to have a lower FPE than BB overall, with the other three methods performing significantly worse. The same pattern was observed with the Fourier basis. Thus, for the Log class there may be a greater dependence on the prior than in the earlier classes.

Log	Bayes			Frequentist	
Method	PWM	BB	BMA	AICMA	STK
FPE Cheb, a=-1, n=100	3.144	3.327	3.464	**2.824**	4.313
SE(FPE)	.313	.408	.028	.023	.028
FPE Sin, a=-1, n=100	**2.462**	2.465	2.904	2.909	3.883
SE(FPE)	.012	.013	.026	.019	.021
FPE Four, a=-1, n=100	**1.267**	1.276	3.56	2.807	3.883
SE(FPE)	.012	.006	.014	.018	.021

Table 7: FPE's and SE's for five methods, three bases, $a = -1$, $n = 100$, and $p = 1$ for the Log function class.

It is interesting that the FPE's tend to decrease from Chebyshev, to Sine, to Fourier, whereas for other function classes the Sine basis tended to do worst. It may be that for Log, the oscillations of Sine are not as harmful because Log rises rapidly at its right end point. This suggests that, even for Log, that PWM is overall a little better than BB.

Table 8 shows our results using the Doppler function class. The differences from one method to another are small. The instability as to which method is best may require many more iterations to resolve, and even so, the differences may remain small. PWM with Fourier performed best; however, across all runs in the Doppler class there was considerable variability in the relative performance of predictors. For the Chebyshev basis, BMA does best, PWM places second, and BB places third for all values of a although they are not formally distinguishable. For the Sine basis and $a = -1$ STK is best for Doppler, but for other values of a AICMA does best with BMA, BB, and PWM being indistinguishable. For the Fourier basis, PWM is best with $a = -1$; however, AICMA again does best for $a \neq -1$.

Since BMA had not done well in previous simulations, we further investigated its performance with the Chebyshev basis. Varying a and b in the generation of Doppler functions made no difference, and neither did reducing the SD of the error distribution, so we studied examples of histograms of the selected models for the Doppler class. They looked qualitatively like the upper left panel in Figure 5. Specifically, the smallest models, consisting of the lowest order terms from the basis, were consistently receiving 85% of the posterior probability. This fact implied that BB and PWM were choosing the same model most of the time, explaining why their errors were nearly equal. This also explained the performance of BMA, namely, it would always be a nontrivial model average putting 15% or so of the posterior probability on models with higher order

Doppler	Bayes			Frequentist	
Method	PWM	BB	BMA	AICMA	STK
FPE Cheb, a=-1, n=100	.830	.831	**.827**	.831	.844
SE(FPE)	.007	.007	.008	.004	.023
FPE Sin, a=-1, n=100	.834	.835	.832	.830	**.824**
SE(FPE)	.004	.004	.004	.005	.007
FPE Four, a=-1, n=100	**.788**	.794	.794	.827	.845
SE(FPE)	.005	.005	.005	.007	.000

Table 8: FPE's and SE's for five methods, three bases, $a = -1$, $n = 100$, and $p = 1$ for the Doppler function class.

terms. Since BMA included higher order terms while and BB and PWM were devolving to model selection, BMA did better.

The question remained as to why the predictors concentrated so heavily on low order terms. This led us to identify the 'Bail out Effect', i.e., when a predictor chooses a very simple model that is quite unlike the complex true model. We conjectured that the symmetry around $y = 0$ may explain the success of BMA. That is, the oscillations of the Doppler function class are so rapid relative to the sample size and bases that they are ignored by the predictor, which 'bails out' to predicting essentially zero because of the symmetry. To test this, we replaced the $\sqrt{x(1-x)}$ factor in Doppler with $1/x$ and $(1 - x)$ but the results did not change as both functions have the same sign on $(0, 1]$. However, when we broke the symmetry by using these versions of Doppler with absolute value on the sine function, we saw the more typical ordering of PWM being slightly better than BB being better than BMA for each basis across all values of a, p and n. This confirms our intuition that when it is predictively good to 'bail out' to a very simple model that is quite unlike the complex true model, BMA does well. That is, when an $\mathcal{M}$-complete problem can be well-approximated by an $\mathcal{M}$-closed problem, then BMA tends to do best.

3.4 Interpretation of Results

Our main point is that PWM and BB do demonstrably better in a predictive sense than BMA, AICMA, and STK. While PWM and BB cannot be separated in terms of SE's, it is almost always the case that PWM has a slightly smaller predictive error than BB taht can be detected as noted after (8).

The exceptions to this general behavior come in two forms. The first is in contexts where a relationship exists between the DG and the inputs to the predictor that a predictive method can exploit. This was seen with the Trig class and Chebyshev basis where the bias-variance tradeoff achieved by STK does well, see Table 2. This was also seen with Log class and the Chebyshev basis where AICMA does well, see Table 7.

However, a second exception is seen in another feature of the Log class that is important.

Because the Log class has a region where the DG is hard to approximate, the Log class may be regarded as closer to the high complexity end of $\mathcal{M}$-complete problems. AICMA does well merely because it permits higher order terms more readily. One can argue that the Doppler class is one step more complex than the Log class and hence even closer to the $\mathcal{M}$-complete-$\mathcal{M}$-open boundary. This may explain the even greater instability of the predictors for the Doppler class: BMA does best with Chebyshev (though poorly, due to the 'Bail-out effect') while AICMA does best (though poorly) for Sine and Fourier. The dependence on the prior also matters with Doppler although we have not discussed this. Indeed, prior selection affects the parameter estimates and can affect which predictor is better, perhaps because prediction for the Doppler class is so unstable.

The results in the earlier tables are surprising from a theoretical standpoint. First, established theory suggests that BMA and BB are roughly comparable (Lee and Oh 2013) for $\mathcal{M}$-closed problems with orthogonal regressors. However, in our simulations with orthogonal regressors in $\mathcal{M}$-complete problems where the degree of approximation should in principle be very good (e.g., Trig and NN's), BMA and BB generally perform very differently. This suggests the problem class – $\mathcal{M}$-closed vs. $\mathcal{M}$-complete – matters more than the possible degree of approximation suggests.

Second, although PWM, BB, and BMA have optimality properties they are conceptually different. First observe that BMA is optimal in a squared error sense and in a logarithmic sense; see Raftery and Zheng (2003). The squared error sense is a posterior risk, i.e., conditional on the data, whereas the logarithmic sense involves an integration over the sample space, i.e., removing the expectation over the data in Theorem 4 of Raftery and Zheng (2003) makes the result false. The optimality of BB is in a posterior sense; see the reasoning in the proof of Theorem 1 in Barbieri and Berger (2004). By contrast, the optimality of PWM, see (2), depends on all the available data including $\mathbf{x}^{\text{new}}$. If the expectation of the summation over k were taken first and then the minimizing action u were found, the result would be the L1-BMA which we conjecture would have similar properties to the L2-BMA. Thus, taking the minimum in the posterior gives a different result when dependence on $\mathbf{x}^{\text{new}}$ is permitted. This may make the PWM more adaptive than the BMA. The same logic implies that the BB is more adaptive than the BMA, but not quite as adaptive as the PWM meaning the PWM is consistently a little better than BB. The reverse may apply to the AICMA and STK: Both are so adaptive to the data that it may harm their performance in $\mathcal{M}$-complete problems – although this might be good for $\mathcal{M}$-open problems. We do not pursue this, but see Leung and Barron (2006) and Clarke (2003).

It is worth remembering that the formal results showing optimality for BMA and BB are in the $\mathcal{M}$-closed case while the simulations here are generally in the $\mathcal{M}$-complete case. Even if BB or BMA outperformed PWM for large enough n in the $\mathcal{M}$-complete case, which seems unlikely, we would want to explain why PWM outperforms BB and BMA (and the other methods) in small samples. To address the difference in performance conceptually, observe that, in addition to representing $\mathcal{M}$-complete DG's, the seven function classes are presented roughly in terms of the difficulty of the predictive task they define, i.e., a sense of complexity. Once the difficulty is high enough, as it begins with the Log and especially Doppler classes, it becomes unclear which method is superior.

Indeed, as seen in Table 8 the various methods perform comparably. We interpret our results as meaning that, aside from exceptional cases, PWM does best on $\mathcal{M}$-complete problems until the difficulty of the problem approaches that of an $\mathcal{M}$-open problem, at which point $\mathcal{M}$-complete methods break down. Likewise, from the variations on the Doppler class that we considered, it seems that methods that are $\mathcal{M}$-closed optimal like BMA become competitive with PWM when the DG approaches the $\mathcal{M}$-closed-$\mathcal{M}$-complete boundary.

A separate issue is the absolute size of the FPE's for different bases. We attribute this not to a complexity interpretation, but rather to a bias-variance analysis. The Sine basis often has elevated FPE's and this may be due to its oscillatory nature. Even if the coefficients of higher order basis elements decrease, the oscillations are roughly twice as fast as for a Fourier basis of the same size (since half the terms are cosines) and the Chebyshev basis does not have oscillations. Moreover, as a generality, the Fourier basis tends to give better predictive performance than either the Chebyshev or Sine, perhaps because the Fourier basis is more complex. So, it is no surprise that the best results are usually achieved with the Fourier basis and PWM. Indeed, one can argue that Fourier is the most complex basis since it has two sorts of terms, sines and cosines. Also, one can argue that PWM is the most robust i.e., resistant to changing predictions unduly as a function of the input data. These factors may explain why PWM with Fourier did overall best.

The last point is that the inclusion of other predictors would likely have little impact on our conclusions. Other possible predictors include (1) the mean of the models' predictions, (2) the median of the models' predictions, (3) a median form of stacking formed by optimizing the median rather than the mean, and (4) the L1-BMA. We did not investigate (1) or (2) because predictors that use the posterior should be better in general. Regarding (3), we did not include a median form of stacking because in simple cases (not shown here) it performed poorly. Finally, the L1-BMA is a nontrivial model average that should be inefficient compared to L2-BMA, and inefficient compared to BB and PWM since they select a model to make predictions. In addition, we expect that L1-BMA would be too stable since it will entirely ignore tail behavior.

There are two other Frequentist techniques that we could have included. First, there is a version of boosting for regression. Basically, the boosting principle is similar to gradient fitting: Greedy fitting iteratively on residuals using weak learners can approximate a true underlying model without too much overfitting. While established as a good technique for classification (despite notable dissent, see Mease and Wyner (2008)), boosting has not done so well for regression. For example, Park et al. (2008) obtains compelling theoretical results but in practice boosted regression rarely out-performs other regression techniques. This corroborates the equivocal findings of Ridgeway et al. (2008) and Schonlau (2005). However, no systematic simulation study has been done. As an intuition, it is unlikely that boosting will help much with regression because boosting techniques for regression effectively turn a regression problem into a very large number of classification problems (by partitioning the domain of the unknown function). Solving these individually by pooling weak learners and then combining the resulting classifiers is probably not going to be as effective as other methods that do not decompose the prediction task into a collection of disjoint tasks.

Second, there is bagging which uses bootstrap samples to improve the stability of a good predictor. However, in each case one must have a base predictor to 'bag'. Thus, the limitation of this approach is that its computational demands were too high to allow a timely and systematic examination. However, individual examples we have done (but do not present here) suggest that bagging AICMA and BMA improves them over their unbagged versions but not enough to outperform the best methods within our comparisons. For some extremely complex real data sets that are probably $\mathcal{M}$-open, bagging or stacking relevance vector machines (that we have not used here) does seem to give some improvement over other methods. We conjecture that when PWM or BB is best, bagging may well make them worse since they are already relatively stable (being based on medians). We did not investigate relevance vector machines or support vector regression because kernel methods seem more appropriate for $\mathcal{M}$-open problems than $\mathcal{M}$-complete problems.

4 Looking Inside PWM and BB

We examine PWM and BB in more detail since they emerged as the best two methods in the simulations of Section 3. We begin by giving several graphs so that the curvefitting approach taken by PWM and BB to generate predictions can be seen. Building on this intuition we then look at the sampling distribution of the models from which PWM and BB give their predictions. This helps us understand bailing out, bailing in, and the 'nice' cases. In the last part of this section, we turn to a localized assessment of variance for PWM, BB, and BMA. We include BMA to understand why it performed so poorly.

4.1 Visualizing PWM and BB

As an aid to understanding the predictive performance of PWM and BB in the previous section, we present several pictures of the curves used by PWM and BB to make point predictions. Figure 3 shows four examples from different function classes, two with $n = 30$ and two with $n = 100$, and two using the Chebyshev basis and two using the Fourier basis since these bases are more common than the Sine basis.

As a generality, the curves found by PWM are a little closer to the true curves. The variability is generally greater when $n = 30$ than when $n = 100$ and the curves with $n = 100$ are closer to each other. In Figure 3, it is seen that the PWM curve is a little rougher than the BB curve is, except for the lower left panel where they appear equally smooth. This arises because PWM may use different models for different ranges of x's unlike BB.

It is interesting to note that the upper left panel with NN's actually has the worst fit (for both PWM and BB) even though PWM fits better than BB over almost the whole range and NN's are very smooth. The upper right plot shows that there is a tradeoff between making accurate predictions between the bumps and making accurate predictions at the

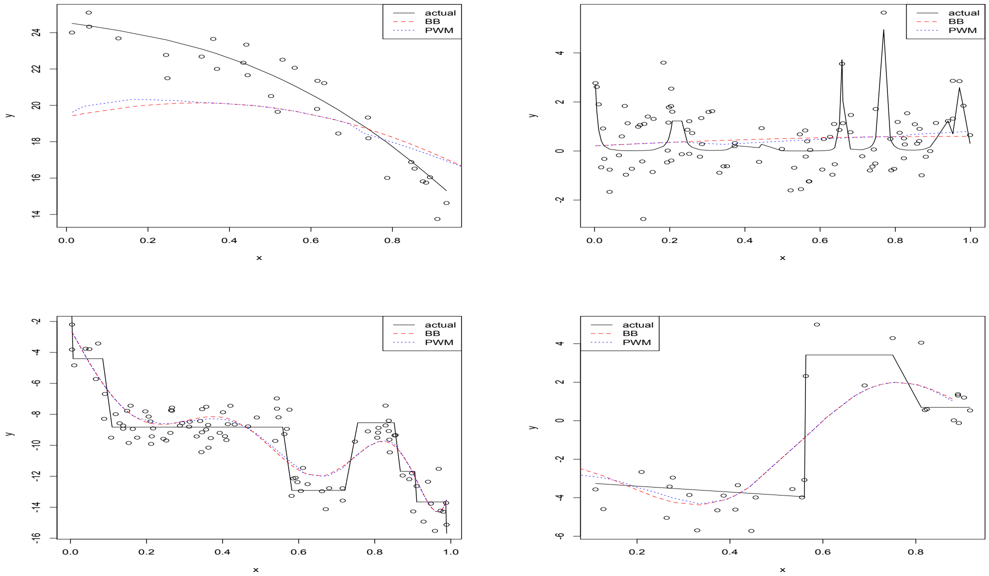

Figure 3: Examples of the curves (solid) fitted by BB (dashed) and PWM (dotted), all cases are $a = -1$. Upper left: NN with the Fourier basis and $n = 30$. Upper right: Bumps with the Chebyshev basis and $n = 100$. Lower left: Block with Chebyshev basis and $n = 100$. Lower right: Treed with the Fourier basis and $n = 30$. In all cases, the open circles indicate the values $y_i = f(x_i) + \epsilon$, where f is the actual function plotted.

bumps. PWM fits the flattish parts of the curves better but does about the same on the bumps themselves, except at the right where PWM is a little higher than BB, although it must be admitted that both methods miss most of the structure in the Bumps class. The lower left plot shows the curves are very close even though PWM is a bit closer to the true function on [.2, .6]. The lower right curve shows that PWM captures the curvature to the left of .4 better than BB does. It must be admitted that these function classes were chosen to represent the mid-range of difficulty of $\mathcal{M}$-complete problems; easier or harder classes might give different interpretations.

Overall, graphs like those in Figure 3 show that PWM predicts better not only because it necessarily fits the curve better (although it usually does) but because the regions where it fits well enable it to give predictions that are on average smaller than the regions where BB fits well. Otherwise put, the regions where PWM fits poorly do not lead, on average, to bigger errors than the regions where BB fits poorly.

4.2 Selection of the Predictors

Because the models are nested, the predictions from BB and PWM can always be regarded as having come from one of the models on the list. This is not the case for BMA, AICMA, or STK. So, in this subsection, we examine the selection of the prediction-giving model for PWM, and for BB. We present histograms estimating the sampling distributions for predictor selection over 2000 uses of PWM and BB, for a fixed basis, a, and n. Each use of PWM and BB draws a function at random from a class and a new set of n x_i's and n error terms to generate the data used to form the PWM and BB predictors. Then, the PWM and BB predictors are evaluated at a new x_{n+1} (drawn from the Unif[0,1], but other distributions would give analogous results) to predict $Y_{n+1}(x_{n+1})$. Thus, we have made 2000 predictions for 2000 x-values, from 2000 predictors for 2000 data sets. The models from which PWM and BB make their predictions are then recorded. Thus, in these histograms, the horizontal axis represents the models in order of size and the vertical axis represents the proportion (as a frequency out of 2000). As before we focus on the cases with $a = -1$ and $n = 100$; however, for the sake of clarity, we occasionally present other cases. Our intuition is that for good prediction a bias-variance tradeoff will lead to histograms that are unimodal around some model in the interior of the model list, i.e., strictly between the smallest one term model and the largest 29 term model. We expect tightly unimodal histograms for smooth functions and loosely unimodal histograms for more complex functions. However, this can fail in two ways – bailing out and bailing in – and these indicate different properties of the prediction problem. So, we present three examples – the 'nice' case, bailing out, and bailing in.

We begin with the 'nice' cases and see they may only be achieved with large enough n. Detailed examination of the histograms reveals that the Blocks, Trig, and Log classes led to qualitatively similar predictor selection behavior, so it is enough to present results for the Log class for $a = 1$ in Figure 4. In this case, the qualitative behavior is essentially independent of a but dependent on the basis and sample size. For the Chebyshev basis, when $n = 30$, the histograms are essentially unimodal at the largest model (this is an

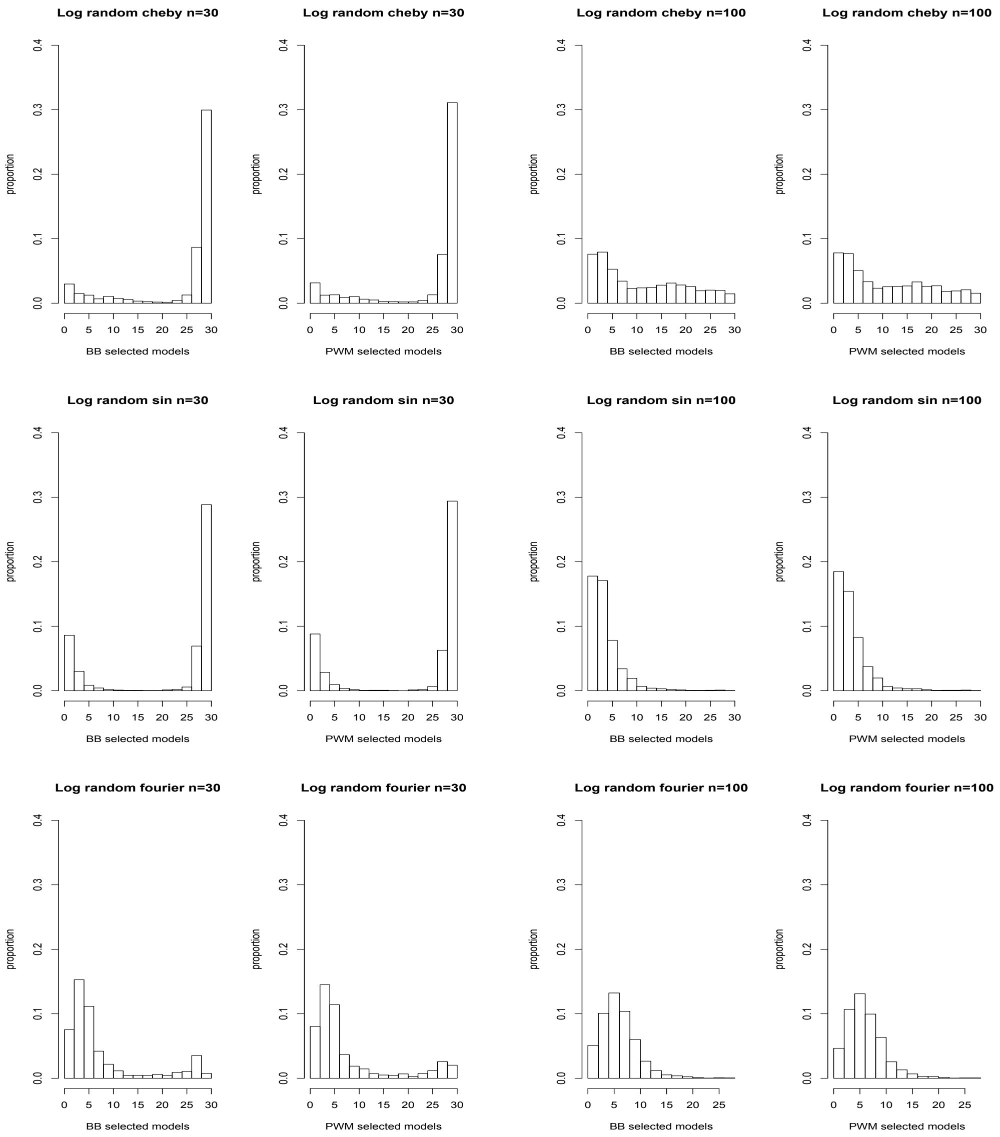

Figure 4: Log class: The rows correspond to bases Chebyshev, Sine, Fourier with $a = 1$. The columns correspond to $n = 30$ on the left and $n = 100$ on the right. The horizontal axis indicates the 29 models in their natural order; the vertical axis indicates the proportion of times a model of each size is used by BB and PWM over the 2000 iterations.

instance of bailing in, see below) but for $n = 100$ the histogram has a very choppy appearance with a heavy right tail and a weak mode around 4. If we increased the sample size, it is unclear how the histogram would behave. Likely, for large enough n the mode would gradually shift to the right. For Sine, when $n = 30$, the histogram is bimodal with a large mode at 29 and a small mode at 1, much like for Chebyshev. When $n = 100$, the mode centers on 3-4 (like Chebyshev) and the right tail is very light (unlike Chebyshev). For the Fourier basis, when $n = 30$, there is a major mode around 4-5 and a small mode around 28. When $n = 100$, the mode has shifted to around 6 and the right tail is quite light. The nice shapes of the panels in the lower right of Figure 4 reveal why PWM and BB with Fourier did better than any other method-basis pair. Loosely, the histogram for the Fourier basis suggests a better variance bias tradeoff than for Chebyshev or Sine.

In the Blocks, Trig, and Log classes only the Fourier basis achieves a nontrivial bias-variance tradeoff, i.e., a nice unimodal histogram, as early as $n = 30$. A strong mode around four/five suggests that smaller models give poor predictions due to bias while larger models give poor predictions due to variance. (The smaller mode at model 29 indicates that when the bias-variance tradeoff is unsuccessful the methods favor the higher variance models.) For Chebyshev and Sine, the larger sample size of 100 is required to begin to see a useful bias-variance tradeoff, even though it remains weak for the Chebyshev basis. As one passes from the Chebyshev to the Sine to the Fourier bases, the histograms look increasingly 'nice', shifting from bimodal or weakly unimodel to a more 'normal' appearance. Note that in Table 7, the Fourier basis gives the best results (with PWM). Chebyshev gives better results (with PWM) for Blocks possibly because of the long right tail in the top right panels, see Table 5, and Chebyshev gives better results with STK for Trig even though PWM is second best (best among the three Bayesian methods), again possibly because of the long right tail, see Table 2.

The histograms for the NN and Treed function classes exhibit different behavior from Trig, Blocks, and Log, but are qualitatively similar to each other. The value of a makes little difference so we present results for $a = -1$. Specifically, for Sine and NN, the histograms in the upper left of Figure 5 put almost all their weight on the smallest model even for $n = 100$. This is what is meant by bailing out. Bailing out may be reflected in the outsized errors of all five methods in Table 3 and to a lesser extent in Table 6. It may mean that no larger model using Sine basis elements achieves a better bias-variance tradeoff.

The upper and lower right panels of Figure 5 show that when $n = 100$ the Fourier and Chebyshev bases have similar histograms, reflecting only slightly different bias-variance tradeoffs. The 'nice' tight shape of the histograms for NN's is reflected in the good performance of PWM and BB in Tables 3 and 6. However, the lower left panel shows that Chebyshev gives a secondary mode at 29 when $n = 30$; no secondary mode is seen in the corresponding histogram for Fourier (not shown). This means Chebyshev may require a larger sample size to give results as good as Fourier for NN. (This is consistent with the FPE's for $n = 30$.) No bailing out is observed in these cases.

We see next that the histograms for the Bumps class indicate a third type of behavior –

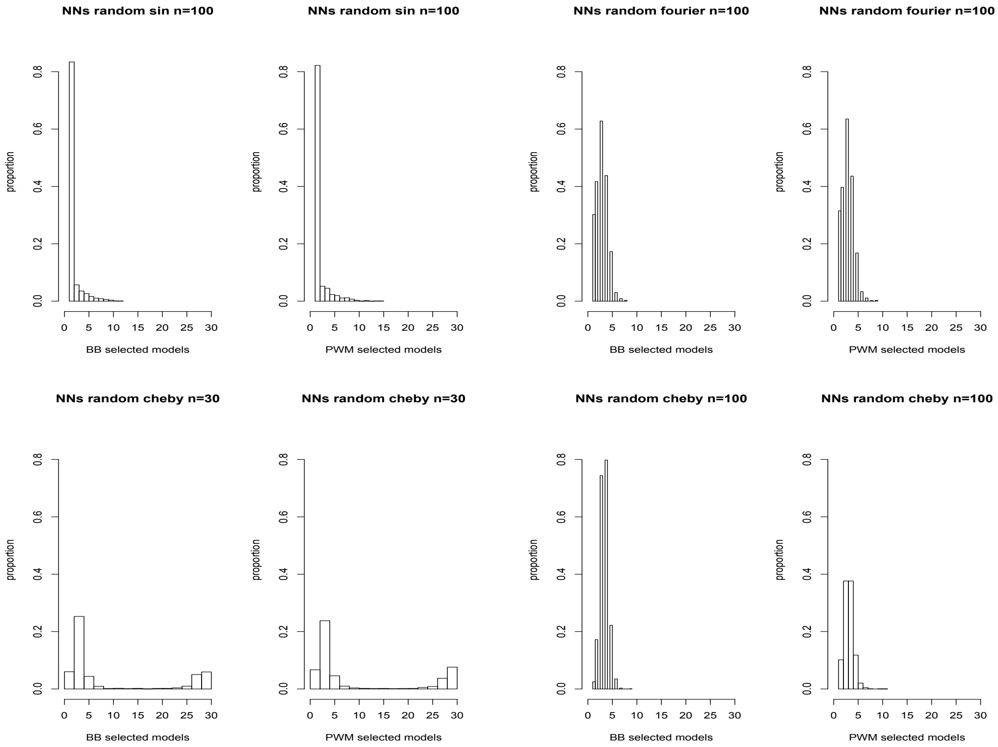

Figure 5: NN class and $a = -1$. Top left panel: results for the Sine basis and $n = 100$. Top right panel: Fourier basis for $n = 100$. Bottom two panels: Chebyshev basis for $n = 30$ and $n = 100$. The horizontal axis indicates the 29 models in their natural order; the vertical axis indicates the proportion of times a model of each size is used by BB and PWM over the 2000 iterations.

bailing in, meaning a convergence to the most complicated model. Bailing in is closely related to bailing out in that bailing in can result from incompletely fixing a bail out problem. For instance, for the Bumps class, in the top row of Figure 6 we see that for $n = 100$, the predictors bail out to a one-term model indicated by a mode at the extreme left hand side and a light tail. The bail out is stronger for $n = 100$ than for $n = 30$. This occurs for all values of a, not just $a = -1$. (The Doppler class shows an even stronger tendency for PWM and BB to bail out.) If we change the prediction problem by increasing the sample size to $n = 350$ and using the Fourier basis, we still observe the bail in for all bases. An instance of this is seen in the bottom left panel of Figure 6. Likewise, if $n = 100$ but we reduce σ to .1 we observe 'bail out' (for all bases).

However, when we both increase n and decrease σ we find different behavior. Fourier and Chebyshev were qualitatively the same, with heavy emphasis on larger models; see the lower right of Figure 6. This is bailing in: Defaulting to the largest possible model. In the same setting, for Sine, we observed results as before: The histograms concentrated at the left indicating bailed out behavior. Thus, even with more information (higher n and lower σ) predictors using the Sine basis still bail out to the smallest model, but predictors using the other two bases bail in to the most complex models. This means that the bias-variance tradeoff for Sine reflects an inability of the predictors to encapsulate the true function while the bias-variance tradeoff for Chebyshev and Fourier favors larger models i.e., tolerates more variance for the sake of reduced bias. Since bailing in is a better choice given the complexity of the DG, we see that Fourier and PWM do best with Chebyshev and PWM second best in terms of FPE's; see Table 4. Note that bailing in is not what we want for predictive purposes: Ideally, we want the mode of the histogram to be in the middle of the model list, not at either endpoint. Since defaulting to the most complex model is often better than defaulting to the simplest model, we suggest bailing in represents an incomplete solution (not enough complexity in the models) to a bail out problem.

Overall, we suggest that the 'nice' cases indicate a good model list has been chosen for prediction as the unimodal shape indicates a meaningful bias-variance tradeoff. However, both bailing out and bailing in mean that the model class is inappropriate. Bailing out means that the complexity of the DG is so high relative to the predictor that the predictor cannot mimic its output, i.e., the predictor defaults to the minimal variance case and gives up on keeping the bias relatively small. In the present examples this indicates the complete inadequacy of the model list. Bailing in is the reverse: The predictor defaults to the minimal bias case and gives up on keeping the variance relatively small. In the present examples this indicates the model list is too small or otherwise unable to provide a parsimonious representation of the DG.

4.3 Variability of FPE as a Function of $\mathbf{x}^{\text{new}}$ for PWM, BB, and BMA

It is not enough just to look at the mean FPE because the variability can depend on the location at which the final prediction is made. So, for each function class we examined the variance of prediction across the input domain, conditional on the other inputs (noise term, a, n). Given $(Y_1, \mathbf{X}_1)$, ..., $(Y_n, \mathbf{X}_n)$ we generated $Y_{n+1}(\mathbf{x}_{n+1})$ and for a

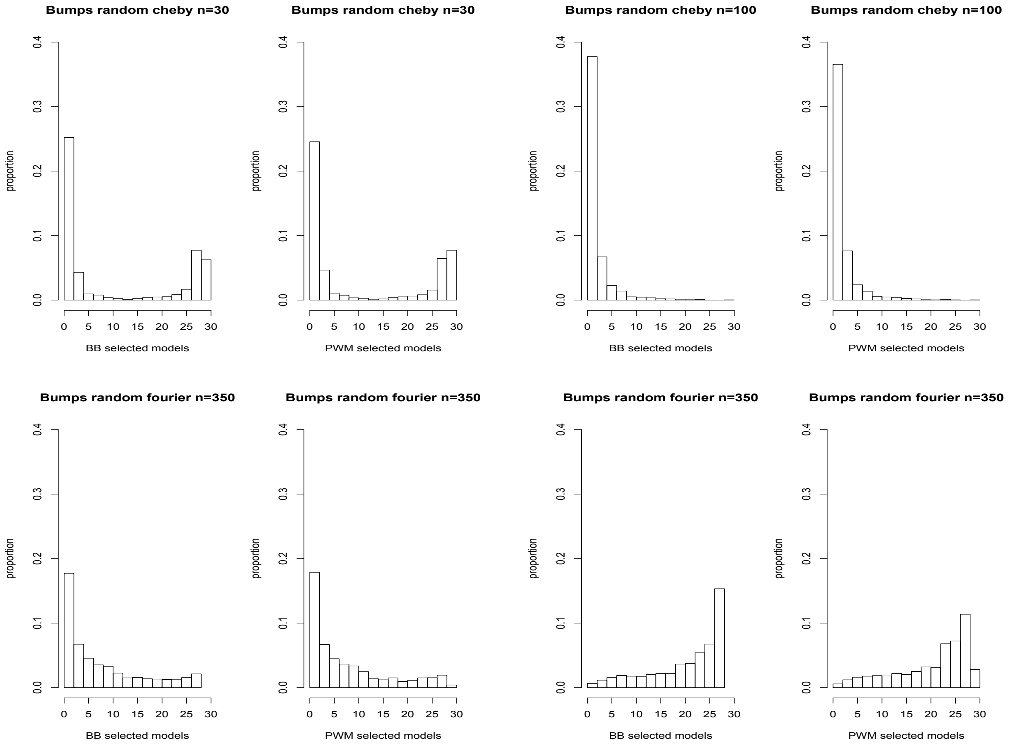

Figure 6: Bumps class and $a = -1$. Top row: Chebyshev basis for $n = 30, 100$. Bottom row: $n = 350$ for Fourier with $\sigma = 1$ (left) and $\sigma = .1$ (right). The horizontal axis indicates the 29 models in their natural order; the vertical axis indicates the proportion of times a model of each size is used by BB and PWM over the 2000 iterations.

given predictor we formed $\hat{Y}(x_{n+1})$. This gives the residual $r_j = |\hat{Y}(x_{n+1}) - Y(x_{n+1})|$. Over m iterations, we therefore get m residuals. Then for each value of $c_g = 0, .05, .1, \ldots$, we found a 'bin' of values $S(c_g) = \{r_j(x_{n+1}) | x_{n+1} \in (c_g - .02, c_g + .02), j = 1, \ldots, m\}$. Since we used 2000 iterations, $\#S(c_1) + \ldots + \#S(c_{20}) \leq 2000$ because a range of x-values of length .01 was omitted between any two successive c_g's. Then, for each g we found the variance of the residuals in $S(c_g)$. This was done for PWM, BB, and BMA, giving a 'varFPE graph' for each function class, basis, n, and value of a. This permits us to localize the 'variability' of a predictor to a range of the domain. The varFPE graphs shown here include all sources of variability except location, i.e., x. Thus, they differ from the SE's presented in Section 3 that average over all the random inputs.

We find that methods that perform poorly tend to have varFPE graphs with large regions where the bin-wise residual variance is high, indicating regions where the bias-variance tradeoff achieved by a predictor is likely poor. This is particularly the case with BMA. As the figures will show there are two typical varFPE graphs, namely, cup-shaped and decreasing to the right. (Decreasing to the left also occurs, but is the mirror of decreasing to the right.) As before when $n = 30$ the graphs were choppier and in some cases suggested an effect from the prior, while when $n = 100$ the prior had little effect and the graphs were smoother.

As an example of a cup-shaped graph we consider the Log class used in Barbieri and Berger (2004). The bottom of the cup indicates a region of x's where the bias-variance tradeoff achieved by a predictor is most successful. As indicated in Figure 7, there is a tendency for the curves to rise rapidly at the right hand endpoint, due to the vertical asymptote of the log function, and to rise at the left, possibly due to an edge effect as the bases are not localized. The Sine basis tends to give higher variability, again indicating that oscillations make achieving a good variance-bias tradeoff more difficult. However, the curve for BMA is below those for PWM and BB for the Sine basis but above those for PWM and BB for the Chebyshev and Fourier bases. Indeed, the curves for BB and PWM are lowest for Fourier, consistent with Table 7. Indeed, the overall poor performance of BMA may be reflected in the rapid rise at the right hand side and lesser rise on the left hand side in the varFPE graphs.

As an example of a decreasing graph we consider the Trig class. In this case we get two distinct behaviors when $n = 100$. For Fourier and Chebyshev, we find a dependence on a but not on n. When $a = 1, 3$, i.e., the prior is diffuse, the graphs for PWM and BB rise on the left; when $a = -1, -3$, the graphs are flat over the whole domain. This is reasonable because sparsity priors will tend to eliminate higher order terms so they contribute less to variability. Representative examples are shown in the two left panels of Figure 8. For Sine the curves had a common shape: The BMA curve was above the essentially coincident BB and PWM curves, and all curves decrease from left to right as seen in the right panel of Figure 8. Thus, BB and PWM perform worse for the Sine basis than for the Chebyshev and Fourier bases; this is seen in the FPE's of Table 2. Indeed, the ordering of the curves from highest to lowest corresponds to decreasing FPE's, as seen in each row of Table 2. Again, the poor performance of the Sine basis may be due to the oscillations of the higher order terms giving a worse bias-variance tradeoff.

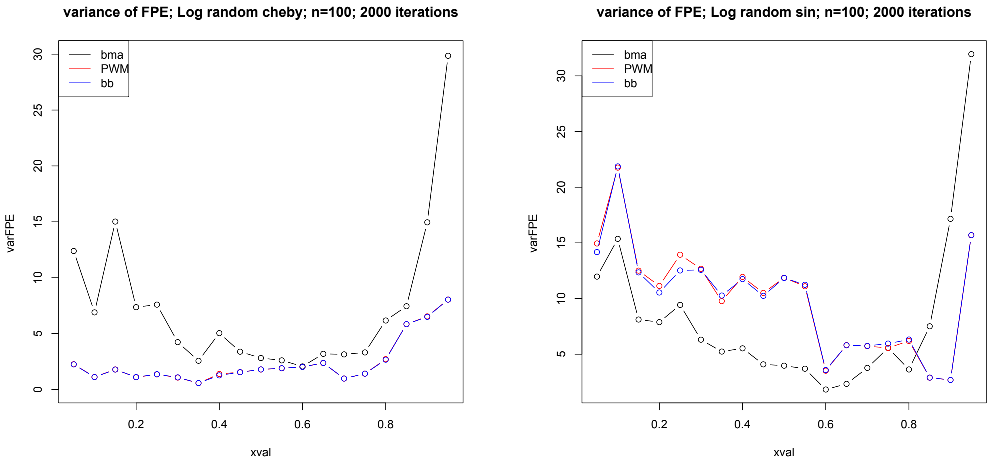

Figure 7: Variance in the FPE for the Log class as a function of x, $n = 100$. Left: Chebyshev, $a = 1$. Right: Sine, $a = 1$. The curves for the Fourier basis with $a = 1$ were similar to those for Chebyshev, but a little lower.

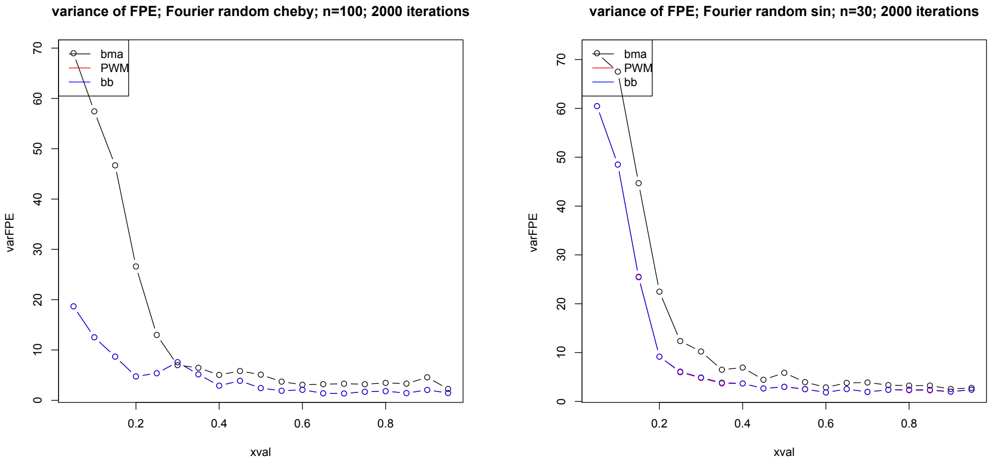

Figure 8: Variance in the FPE for the Trig class as a function of x. Left: Chebyshev, $a = 3$ and $n = 100$. The graph for Chebyshev with $a = -33$ and $n = 100$ is similar but the curves for BB and PWM are flat and near zero over their whole domain. Right: Sine, $a = 3$, $n = 30$.

We comment that the corresponding graphs for the Bumps, Block and Treed cases exhibited a U-shape, somewhat like Figure 7, but without the rise at the left hand side. Also, the graphs for Doppler and NN's were too choppy to be informative apart from showing that BMA was more variable than either BB or PWM.

In aggregate, we have seen that BMA tends to have a higher variance and therefore probably a worse bias-variance tradeoff than PWM or BB, possibly because BMA includes all models without selection. The varFPE graphs support the view that PWM with Fourier tends to do best in a bias-variance sense and hence in a predictive error sense. The method-basis pairings PWM-Chebyshev and BB-Fourier tend to do second best. Note also that decreasng (or increasing) varFPE graphs can be regarded as the left (or right) sides of a cup shaped varFPE graph.

5 Real Data: Higher Dimensional Examples

In this section we apply the five predictive techniques to two $\mathcal{M}$-complete data sets, the second more complex than the first, for which exact modeling is unlikely to be effective. From the earlier sections, we expect that (i) the prior will make little difference for sample sizes approaching 100 or larger, (ii) PWM will do a little better than BB and both will outperform BMA and the Frequentist methods, (iii) the Fourier basis will do better than the Chebyshev, which will outperform Sine, (iv) for the more complex of the two data sets, the identification of the best method-basis pair will be like the Doppler example, i.e., not very clear, and (v) STK and AICMA will not do well outside relatively few special cases. These expectations are largely but not entirely borne out in the FPE tables and in the histograms for term selection presented below. In particular, Sine does unexpectedly well in the first data set (unlike (iv)) perhaps because the data is more oscillatory than we thought. (We do not give varFPE graphs in these examples because they are multidimensional in X, making the varFPE graphs for the individual explanatory variables hard to interpret.)

5.1 Concrete Data

We consider the data set Concrete Compressive Strength publicly available from the UC Irvine Machine Learning Repository. The sample size is $n = 1030$ and there are eight explanatory variables related to the way concrete is manufactured denoted X_1 through X_8. The dependent variable Y is the compressive strength. Details and references can be found at http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength .

To begin, we set the basis elements from which we construct models. Since it is quite hard to work with the Chebyshev basis in eight dimensions, we chose instead a Polynomial (poly) basis. The nonorthogonality of these basis elements might be a problem in other examples, but here the results are in line with the previous simulations with the orthogonal Chebyshev basis. We chose all terms of second order or less among the eight covariates, i.e., all linear terms $X_1, \ldots, X_8$, all squares $X_1^2, \ldots, X_8^2$, and all distinct rectangular terms $X_j X_k$ for $j < k$, for a total of 44 terms. We chose the same number of terms for the Sine basis by taking Sine of the 44 terms from the polynomial basis but using a factor of two in place of a square, i.e., $\sin X_1, \ldots, \sin X_8$, $\sin 2X_1, \ldots, \sin 2X_8$ and $\sin X_j X_k$ for $j < k$. For the Fourier basis we mirrored the polynomial and Sine basis by starting with $\sin X_1, \ldots, \sin X_8$, $\cos X_1, \ldots, \cos X_8$, $\sin 2X_1, \ldots, \sin 2X_8$, and $\cos 2X_1, \ldots, \cos 2X_8$, giving 32 terms. For the remaining 12 terms we chose the best six of the $\cos X_j X_k$'s and the best six of the $\sin X_j X_k$'s based on the correlation with the first hundred observations within each run.

This level of adaptivity may appear to give the Fourier basis a 'head start' over the other bases but it is a reasonable way to ensure all model lists have the same size. We also accounted for this by treating the prediction based on the 101^{st} - 210^{th} observations under Fourier as de facto comparable with prediction after the 1^{st} - 310^{th} observations with Polynomial or Sine. We limited attention to the cases $a = \pm 1$ since they would be indicative of performance for a mildly sparse prior ($a = -1$) or a mildly diffuse prior ($a = 1$). We continued using the uniform prior over the models using the 44 terms.

FPE's

In Table 9 we present the FPE's for PWM, BB, BMA, AICMA and STK at three stages of data accumulation - 30% ($n = 310$), 70% ($n = 720$), and 100% ($n = 1030$). First, note that the numbers are systematically largest for poly and smallest for Fourier, with Sine in between. Second, within each ensemble, the errors decrease as the sample size increases; this is what we expect. Third, the FPE's for the sparsity prior $a = -1$ are lower for Sine and Fourier bases than for the diffuse prior $a = 1$, but higher for poly with $a = -1$ than for $a = 1$. We show the case $a = 1$ for poly and the case $a = -1$ for Sine and Fourier. Fourth, as we expect, the errors increase as p increases.

PWM with Fourier gave the best predictive performance, and BB with Fourier was a close second. This is consistent with the simulations that showed BB is usually a close second to PWM when given the same inputs. As can be seen from the increase in FPE's, the third and fourth best methods - PWM with Sine or BB with Sine (although for $p = 2$ and Sine, AICMA did best) - were not competitive. In fact, Table 9 shows that over the scenarios the methods group into three classes: the best two (PWM and BB with Fourier) with FPE's between 15 to 18 (for $p = 1$), the worst two (PWM and BB with poly) with FPE's between 64 and 89 (for $p = 1$), and the rest with FPE's between 36 and 37 (for $p = 1$). As a curiosity, we have included an extra column labelled MID to represent the predictor taking the middle value as in $\underset{k=1,\ldots,K}{\text{median}}[w_k f_k(\mathbf{x}^{\text{new}}|\tilde{\beta}_k)]$. Although lacking any formal justification, MID did best for the Polynomial basis, possibly because all the w_k's are less than one providing some shrinkage.

Model Selection for PWM and BB

Here we examine the best two methods, PWM and BB, by looking at the selection of models formed from the ensemble. First, recall from Table 9 that the FPE's for the

	PWM	BB	BMA	MID	AICMA	STK
Poly	88.744	88.749	36.442	**36.13**	36.636	35.863
$n = 310$	1107.93	1107.96	237.892	**235.724**	239.598	232.044
$a = 1$	15099.25	15099.42	1614.77	**1601.34**	1628.975	1560.935
$n = 720$	69.449	69.449	36.209	**35.749**	36.209	35.747
	699.342	699.342	234.92	**231.39**	234.92	231.399
	7609.837	7609.837	1584.722	**1559.469**	1584.722	1558.127
$n = 1030$	64.362	64.362	36.137	**35.628**	36.137	35.67
	606.086	606.086	232.968	**229.075**	232.968	230.425
	6046.485	6046.485	1557.481	**1529.664**	1557.481	1548.028
Sine	**35.807**	35.815	36.112	36.13	36.087	35.863
$n = 310$	239.416	239.503	235.542	235.724	235.296	**232.044**
$a = -1$	1680.176	1680.977	1599.719	1601.341	1597.407	**1560.935**
$n = 720$	**34.586**	**34.586**	35.719	35.749	35.677	35.747
	226.482	**226.482**	231.091	231.39	230.695	231.399
	1552.664	**1552.664**	1556.723	1559.469	1553.27	1558.127
$n = 1030$	**34.583**	**34.583**	35.599	35.628	35.559	35.669
	225.659	**225.659**	228.793	229.075	228.42	230.425
	1541.65	1541.65	1527.185	1529.664	**1523.927**	1548.028
Fourier	**17.577**	17.591	35.822	36.493	35.79	35.575
$n = 210$	**89.209**	89.308	232.226	238.374	231.944	229.949
$a = -1$	**490.167**	490.811	1565.76	1618.74	1563.333	1548.389
$n = 620$	**15.99**	15.996	36.123	36.824	36.119	35.217
	76.942	76.977	234.828	241.256	234.773	225.911
	400.9	401.05	1587.71	1643.24	1587.115	1507.914
$n = 930$	**15.651**	15.652	35.138	35.828	35.136	35.164
	75.598	75.621	225.986	232.215	225.962	226.259
	396.457	396.644	1512.814	1565.875	1512.526	1517.933

Table 9: FPE's for Concrete for the three bases, two values of a, three loss functions and various sample sizes. The three rows for each sample size and basis correspond to $p = 1, 1.5, 2$. For Fourier the FPE's represent the prediction of 210, 620, and 930 as 100 data points were used for term selection.

poly basis were systematically higher than for all other cases. Figure 9 reflects this by showing that PWM and BB bail in to the whole model list, i.e., the large errors reflect the inability of either method to use the variables to mimic the response or even discriminate well over the explanatory variables. Note that none of the other methods, BMA, AICMA, and STK, performed as badly.

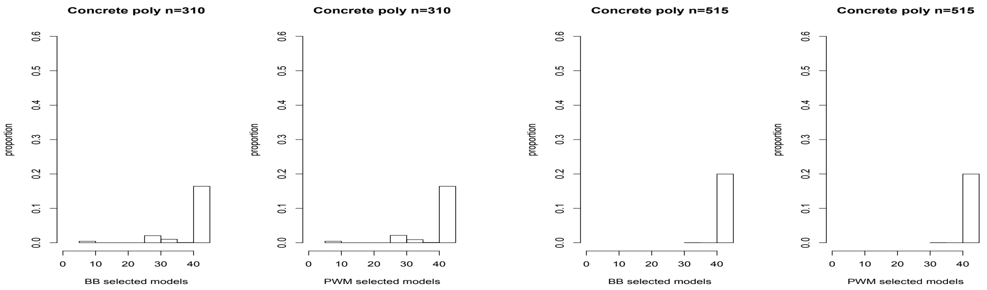

Figure 9: BB and PWM model selection proportions for the Concrete data using the poly basis with $a = 1$. On the left $n = 310$, on the right $n = 515$. The horizontal axis indicates the 29 models in their natural order; the vertical axis indicates the proportion of times a model of each size is used by BB and PWM

By contrast, Figure 10 shows that with the Sine basis BB and PWM (and BMA) put all mass on a single model, model 17, consisting of $\sin X_1, \ldots, \sin X_8$, $\sin 2X_1, \ldots, \sin 2X_8$ and $\sin X_1 X_2$. In the limit, all five methods appear to default to model selection as opposed to a nontrivial model average. This explains why the FPE's with Sine are in such a narrow range.

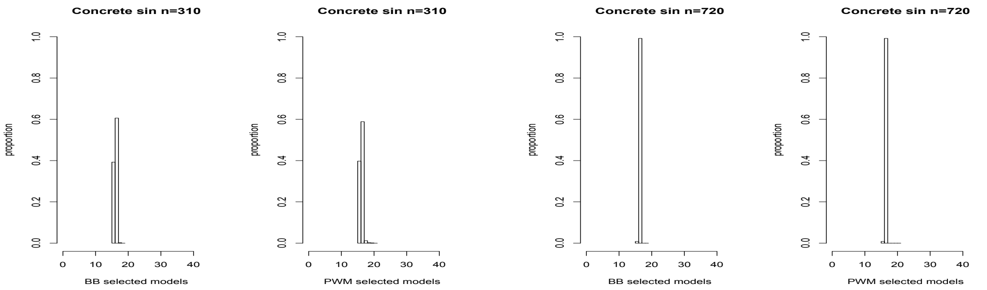

Figure 10: BB and PWM model selection proportions for the Concrete data using the Sine basis with $a = 1$. Left: $n = 310$. Right: $n = 720$. The horizontal axis indicates the 29 models in their natural order; the vertical axis indicates the proportion of times a model of each size is used by BB and PWM.

PWM and BB with the Fourier basis give the best results and the reason is seen in Figure 11. Although we do not observe a 'nice' unimodal histogram at any stage of data accumulation, what we do see is a march toward richer models as n increases. It appears that the predictors choose larger and larger models because the sample size can support it. Throughout, nontrivial averaging takes place with all three methods, i.e., none of them default to model selection. This adaptivity is what may be making PWM with Fourier (and $a = -1$) the best of the prediction strategies. That is, even though we do not obtain a 'nice' unimodal histogram (except in a very approximate sense) as was found in the well-behaved simulations, the model list is large enough to be predictively useful even if it cannot provide a function close to the DG.

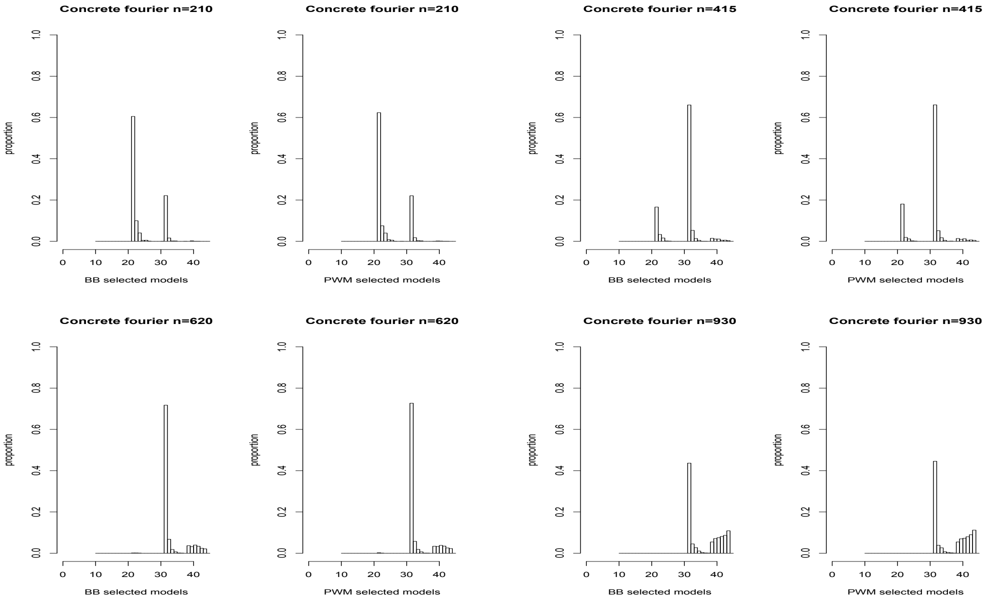

Figure 11: BB and PWM model selection proportions for the Concrete data using the Fourier basis with $a = 1$. Upper left: $n = 210$. Upper right: $n = 415$. Lower left: $n = 620$. Lower right: $n = 930$. The horizontal axis indicates the 29 models in their natural order; the vertical axis indicates the proportion of times a model of each size is used by BB and PWM.

5.2 Forest Fires **Data**

As a second and more complex $\mathcal{M}$-complete example, consider the Forest Fires data set publicly available from the UC Irvine Machine Learning Repository. The sample size is $n = 517$ and there are 12 explanatory variables related to the severity of a forest fire. The 11^{th} variable is virtually constant so we dropped it from our analysis, relabeling the remaining variables X_1 through X_{11}. The dependent variable, Y, is the burn area of the forest fire. Details and references can be found at http://archive.ics.uci.edu/ml/datasets/Forest+Fires .

As before, we set the basis elements from which we construct models. Defaulting to a poly basis instead of a multidimensional Chebyshev basis, we chose 77 explanatory variables: the 11 linear terms $X_1, \ldots, X_{11}$, the 11 square terms $X_1^2, \ldots, X_{11}^2$ plus all the rectangular terms $X_j X_k$ for $j \neq k$. For the Sine basis, we chose $\sin X_j$, $\sin 2X_j$ for $j = 1, \ldots, 11$ and $\sin X_j X_k$ for $j \neq k$. For the Fourier basis we started with $\sin X_j$, $\cos X_j$, $\sin 2X_j$ and $\cos 2X_j$ for $j = 1, \ldots, 11$ giving 44 terms. To find 33 more terms we used the 17 terms from $\sin X_j X_k$ $j \neq k$ and the 16 terms from $\cos X_j X_k$ $j \neq k$ with the highest absolute correlation with the response for the first 100 data points within each run. Again, we limited attention to $a = \pm 1$ and used a uniform prior over models.

FPE's

In Table 10 we present the FPE's for PWM, BB, BMA, AICMA, and STK at three stages of data accumulation - 25% ($n = 129$), 50% ($n = 259$), and 100% ($n = 517$). As before we have adjusted the Fourier sample sizes to reflect the 100 data point burn in but we also used prediction at $n = 189$, as $n = 129$ left an insufficient amount of data (only 29 points) for model estimation and selection.

Table 10 is different from Table 9 in seven ways. First, BMA and AICMA are almost always best when $p = 1$ even though they are optimal under L^2. Second, one of BB and PWM always wins when $p = 1.5, 2$ even though medians are commonly associated with $p = 1$. Third, the FPE's do not always decrease as n increases except for PWM and BB; for BMA, AICMA, and STK, they often increase before decreasing, e.g., STK with each basis, or just increase, e.g., BMA and Sine. Fourth, although not shown, sparsity or diffuseness in the prior makes essentially no difference for poly while sparsity actually harms prediction overall for both Sine and Fourier. Fifth, the polynomial basis frequently gives the smallest FPE's while Sine and Fourier are roughly equivalent. Sixth, for each p, the difference in performance across predictors seems small. Seventh, there are scenarios in which each method is best but no obvious pattern in performance can be seen. Six and seven are reminiscent of the results for the Doppler class in Section 3, consistent with the notion that Fires is more complex than Concrete. The natural conclusion is that the DG is just too complex for the predictors to be effective; this is borne out in the next subsection.

Model Selection

Again, we examine model selection for PWM and BB since they are the only methods that make predictions from a specific model. Given the appearance of Table 10 the model selection probabilities are not entirely unexpected; clearly something is amiss. In fact, for $n > 259$, the model selection probabilities were only nonzero for a very narrow range of functions of very small models. As can be seen in Figure 12, with the poly basis the methods essentially selected the second model while with the Sine and Fourier bases the methods selected the third or fourth model for both $a = 1$ and $a = -1$. That is, the inference from the predictors is that the model containing X_1 and X_2 is best when using the polynomial basis while a model using X_1, X_2, X_3, and X_4 is best when

$a = 1$	L^p	PWM	BB	BMA	AICMA	STK
Poly	1	1.147	1.149	**1.086**	**1.086**	1.088
$n = 129$	1.5	**1.451**	1.454	1.778	1.777	1.767
	2	**2.014**	2.019	3.110	3.108	3.055
$n = 259$	1	1.153	1.153	1.117	1.117	**1.105**
	1.5	**1.458**	**1.458**	1.839	1.839	1.779
	2	2.027	**2.026**	3.237	3.237	3.043
$n = 517$	1	1.142	1.142	1.130	1.13	**1.069**
	1.5	**1.432**	**1.432**	1.950	1.85	1.731
	2	**1.965**	**1.965**	3.227	3.227	2.975
Sine	1	1.169	1.173	**1.086**	**1.086**	1.088
$n = 129$	1.5	**1.581**	1.588	1.782	1.781	1.767
	2	**2.387**	2.402	3.119	3.117	3.055
$n = 259$	1	1.159	1.159	1.117	1.118	**1.105**
	1.5	1.576	**1.575**	1.843	1.843	1.779
	2	2.407	**2.405**	3.246	3.245	3.043
$n = 517$	1	1.160	1.160	1.131	1.131	**1.069**
	1.5	**1.542**	**1.542**	1.854	1.853	1.731
	2	**2.290**	**2.290**	3.235	3.233	2.975
Fourier	1	1.164	1.167	**1.113**	**1.113**	1.131
$n = 89$	1.5	**1.582**	1.589	1.805	1.803	1.859
	2	**2.392**	2.406	3.106	3.102	3.262
$n = 159$	1	1.178	1.179	1.123	1.123	**1.071**
	1.5	**1.589**	1.591	1.840	1.839	1.748
	2	**2.389**	2.393	3.210	3.208	3.035
$n = 417$	1	1.157	1.157	**1.087**	**1.087**	1.127
	1.5	**1.533**	**1.533**	1.793	1.792	1.854
	2	**2.268**	**2.268**	3.150	3.148	3.228

Table 10: FPE's for Fires for the three bases, $a = 1$, three loss functions and various sample sizes. Note that the sample size is adjusted for the Fourier case to allow for term selection. We did not include the MID predictor because it peformed poorly.

using the Sine basis or Fourier basis. As indicated in Figure 12, the convergence to a single model is essentially complete for all bases at some stage between $n = 189$ and $n = 259$. However, these simple models are hardly 'true', and their predictions are relatively poor.

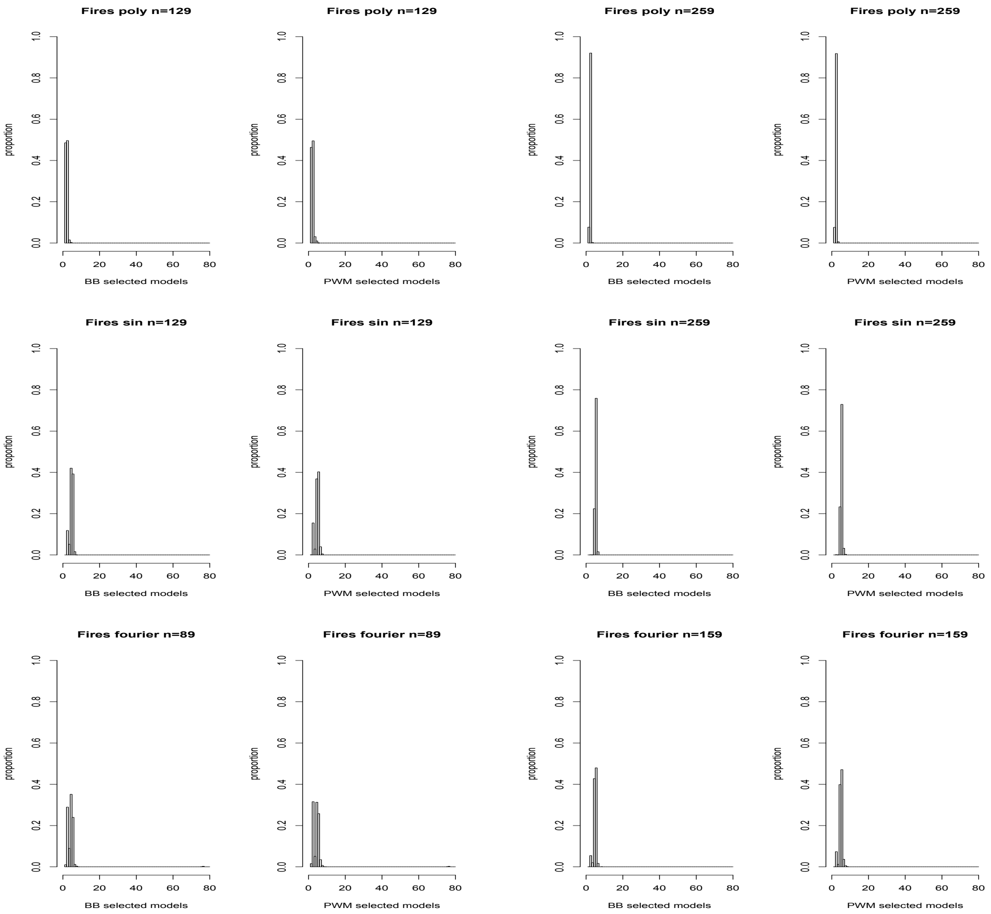

Figure 12: BB and PWM model selection proportions for the Fires data with $a = 1$. Top: Polynomial basis. Left $n = 129$, right $n = 259$. Middle: Sine basis. Left $n = 129$, right $n = 259$. Bottom: Fourier basis. Left $n = 89$, right $n = 159$. The horizontal axis indicates the 29 models in their natural order; the vertical axis indicates the proportion of times a model of each size is used by BB and PWM.

Taken together, the lack of any strong pattern in Table 10 and the high degree of concentration of posterior probability on specific small models points to a version of the bail out effect. In contrast to the Concrete data set, the model list is not large enough to be predictively useful or to provide a function close to the DG.

We tested this conjecture by clustering the data using the terms in the selected models. If the data represent a single cluster this is evidence against the bail out effect; otherwise this is circumstantial evidence for the bail out effect. We used partitioning around medoids clustering on the four-dimensional data $(Y_i, X_{2,i}, X_{3,i}, X_{4,i})$ with two optimality criteria. This is implemented in the R package fpc (function pamk) (Hennig 2013). The first criterion was the Calinski-Harabasz index based on the ratio of the between-cluster means and the within-cluster variances. The second criterion was the average silhouette width. Built into the software is the Duda-Hart test for whether one cluster suffices.

The optimal clusterings under the two criteria are given in Figure 13. Using average silhouette width two clusters are found to be optimal, while under Calinski-Harabasz we find that five clusters are optimal. We conclude that the Fires data may be the mixture of two or more populations. This suggests that the key features of this data set cannot be readily summarized by the available model lists. If the data is an IID mixture of populations then the structure of our predictors, treating the data as exchangeable, is incorrect and a refinement of the procedure, e.g., partitioning the data into its sub-populations, might be necessary. Indeed, one is led to suspect that while Fires is an $\mathcal{M}$-complete problem, it is near the boundary with $\mathcal{M}$-open problems.

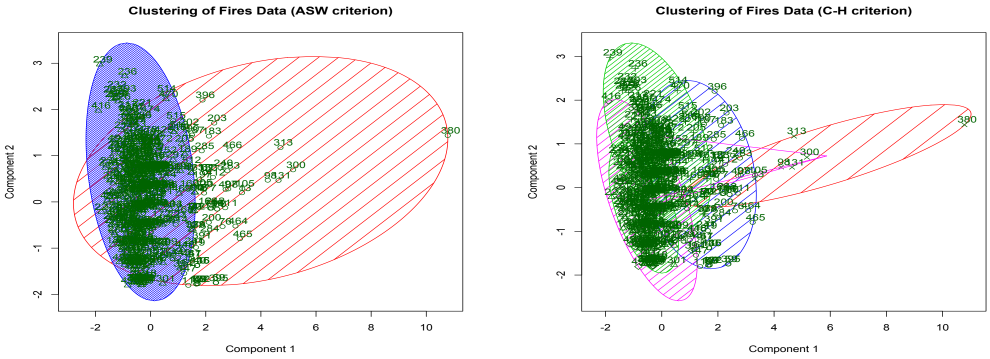

Figure 13: Two clusterings of the Fires data under different optimality criteria for choosing the number of clusters. In both cases, the graphs are planes in the first two principal components (labeled component 1 and 2) which accounted for 59% of the variability in the data. Left: average silhouette width (ASW). Right: Calinski-Harabasz index (C-H). The numbers indicate data points.

6 Discussion

We have proposed a new technique PWM for Bayes prediction for the $\mathcal{M}$-complete class of problems and verified that it works better than other predictors for a large collection of simulations. Admittedly, of the other predictors only AICMA was intended for $\mathcal{M}$-complete problems; BMA, BB, and STK were intended for $\mathcal{M}$-closed problems even though they are commonly used in $\mathcal{M}$-complete settings. In particular, we have used three different classes of ensembles based on different bases. Within each predictive problem there is a bias-variance interpretation, while, in aggregate, the patterns of performance may be better explained by complexity considerations.

We posit that there is range of difficulty associated with the class of $\mathcal{M}$-complete problems. The simulated examples in this paper have been ordered roughly in terms of increasing complexity of the DG as have the real data examples. Our new method seems to work well, outside of special cases, on $\mathcal{M}$-complete problems until the complexity of the $\mathcal{M}$-complete problem approaches that of an $\mathcal{M}$-open problem or reduces to that of an $\mathcal{M}$-closed problem. Indeed, our results for very complex $\mathcal{M}$-complete problems suggest that methods for $\mathcal{M}$-complete problems break down for $\mathcal{M}$-open problems. (Methods for $\mathcal{M}$-open problems likely will do poorly for $\mathcal{M}$-complete problems since they can not use the information that a model exists.) One of the implications of this is that adaptive methods, like PWM, will do better in limited sample size settings than other less adaptive methods. Wong and Clarke (2004) is an example of this in the $\mathcal{M}$-closed case. Adaptivity in the $\mathcal{M}$-open case does not seem to have been explored, but it is not hard to imagine adaptivity will be helpful there, too.

On the topic of model list selection we have seen that good prediction is associated with histograms for model selection that are 'nice', i.e., unimodal, light-ish tails, roughly symmetric. Thus, when the histogram concentrates at the simplest models we are led to a 'bail out effect' in which the predictor scheme gives up on controlling bias and just chooses a predictor with the smallest possible variance. On the other hand, when the histogram concentrates on the most complex models we are led to a 'bail in effect' meaning that the bias is being reduced so much that controlling the variance is not possible. This is another version of the bias-variance tradeoff on the level of model lists found in Fokoue and Clarke (2011). When the predictor bails out or bails in, we are led to rechoose the model list so that models giving serviceable predictions will be in the middle region of the list. One way to do this was used in Clarke and Clarke (2009); however, it is infeasible for limited sample sizes.

In this paper we have not explored model list reselection. Because of the sequential nature of our work we have implicitly permitted model reselection from time step to time step but from the same ensemble. Tackling the more complex predictive problems in the $\mathcal{M}$-complete class of DG's may require periodic reselection of the model list so as to move it – or at least some of its members – closer to better predictors. As a generality, the collection of models is so vast that merely choosing a large collection of models will require more data than will generally be available. Thus, a search over model lists, as well as over models on a list, will have to be done efficiently.

We conclude that prediction is likely to be unsuccessful when the predictive difficulty of the DG is too high relative to the predictor. The concept of the predictive difficulty of a DG is one of complexity and it is reasonable to imagine that there is an analogous concept of complexity appropriate for predictors. Indeed, the bail in or bail out effects may represent a mismatch of the complexity of the DG and predictor as much as a bias-variance phenomenon. In these terms, our prescription that the histogram of model selection be 'nice' can be regarded as complexity matching: For optimal performance the complexity of the predictor should match the complexity of the DG. The impediment is how to formalize the complexity of a DG and the complexity of a predictor so that they are useful and comparable. Outside $\mathcal{M}$-open problems, codelength is one obvious approach. A codelength can easily be defined for a DG, e.g., via the empirical entropy or via two stage codes, see Barron and Cover (1991). Also, a codelength can easily be defined for some predictors such as the BMA (merely code the models, priors, and data with respect to classes of each). On the other hand, the number of terms in a model provides a sort of complexity even if it does not include any assessment of the complexity of the inputs to a predictor or the complexity with which a predictor uses its inputs. That there should be a meaningful notion of complexity for predictors is not controversial. Identifying such a notion formally, however, will be.

References

Barbieri, M. and Berger, J. (2004). "Predictive model selection." *Annals of Statistics*, 32: 870–897. 649, 652, 654, 656, 658, 666, 676

Barron, A. and Cover, T. (1991). "Minimum complexity density estimation." *IEEE Transactions on Information Theory*, 37: 1034–1054. 688

Bernardo, J. and Smith, A. (2000). *Bayesian Theory*. Chichester: John Wiley and Sons. 647, 648, 655

Breiman, L. (1996). "Stacked generalizations." *Machine Learning*, 24: 49–64. 654

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge: Cambridge University Press. 648, 649

Clarke, B. (2003). "Comparing Bayes model averaging and stacking when model approximation error cannot be ignored." *Journal of Machine. Learning Research*, 4: 683–712. 648, 654, 661, 666

Clarke, B. and Clarke, J. (2009). "Prequential analysis of complex data with adaptive model reselection." *Statistical Analysis and Data Mining*, 2: 274–290. 687

Clarke, B., Clarke, J., and Yu, C.-W. (2014). "Statistical problem classes and their links to information theory." To appear in *Econometric Reviews*. 648, 653, 658

Clyde, M. (2012). "Bayesian perspectives on combining models." Slides from presentation at ISBA Kyoto (personal communication). 648, 649

Donoho, D. and Johnstone, I. (1994). "Ideal spatial adaptation by wavelet shrinkage." *Biometrika*, 425–455. 655, 656

Fokoue, E. and Clarke, B. (2011). "Variance bias tradeoff for prequential model list selection." *Statistical Papers*, 52: 813–833. 687

Hennig, C. (2013). "fpc: Flexible procedures for clustering."
URL http://cran.r-project.org/web/packages/fpc/index.html 686

Hoeting, J., Madigan, D., Raftery, A., and Volinksy, C. (1999). "Bayes model averaging: A tutorial." *Statistical Science*, 14: 382–417. 649, 652

Johnson, J. and Omland, K. (2004). "Model selection in ecology and evolution." *Trends in Ecology and Evolution*, 19: 101–108. 649, 653

Kashyap, R. (1980). "Inconsistency of the AIC rule for estimating the order of autoregresive models." *IEEE Transactions on Automatic Control*, 25: 996–998. 653

Lee, J. and Oh, H. (2013). "Bayesian regression based on principal components for high dimensional data." *Journal of Multivariate Analysis*, 117: 175–192. 666

Leung, G. and Barron, A. (2006). "Information theory and mixing least squares regressions." *IEEE Transactions on Information Theory*, 52: 3396–3410. 653, 666

Mease, D. and Wyner, A. (2008). "Evidence contrary to the statistical view of boosting." *Journal of Machine Learning Research*, 9: 131–156. 667

Park, B., Lee, Y., and Ha, S. (2008). "On boosting kernel regression." *Journal of Statistical Planning and Inference*, 138: 2483–2498. 667

Raftery, A. and Zheng, Y. (2003). "Discussion: Performance of Bayes model averaging." *Journal of the American Statistical Association*, 98: 931–938. 648, 666

Ridgeway, G., Madigan, D., and Richardson, T. (2008). "Boosting methodology for regression problems."
URL https://sites.google.com/site/gregridgeway/papers-and-software 667

Schonlau, M. (2005). "Boosted regression: An introductory tutorial and a Stata plugin." *Stata*, 5: 330–354. 667

Shao, J. (1997). "An asymptotic theory for model selection." *Statistica Sinica*, 7: 221–264. 654

Shibata, R. (1981). "An optimal selection of regression parameters." *Biometrika*, 68: 45–54. 653

— (1983). "Asymptotic mean efficiency of a selection of regression variables." *Annals of the Institute of Statistical Mathematics*, 35: 415–423. 653

Shtarkov, Y. (1987). "Universal sequential coding of single messages." *Problems in Information Transmission*, 23: 3–17. 648, 649

Skouras, K. and Dawid, P. (1998). "On efficient point prediction systems." *Journal of the Royal Statistical Society, Series B*, 60: 765–780. 648

Smyth, P. and Wolpert, D. (1999). "Linearly combining density estimators via stacking." *Machine Learning*, 36: 59–83. 649

Symonds, M. and Moussalli, A. (2010). "A brief guide to model selection, multimodel inference and model averaging in behavioral ecology using Akaike's information criterion." *Behavioral Ecology and Sociobiology*, 65: 13–21. 653

Wolpert, D. (1992). "On the connection between in-sample testing and generalization error." *Complex Systems*, 6: 47–94. 648, 649, 654

Wong, H. and Clarke, B. (2004). "Improvement over Bayes prediction in small samples in the presence of model uncertainty." *Canadian Journal of Statistics*, 32: 269–283. 649, 652, 687