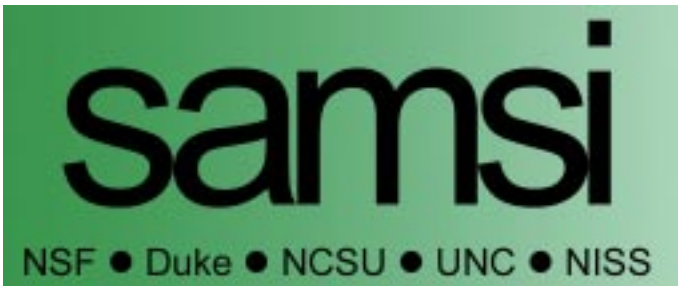


Optimal Model List Selection for Prediction

Ernest Fokoue and Bertrand Clarke

Technical Report #2004-20
August 27, 2004

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Optimal Model List Selection for Prediction

Ernest Fokoué* **Bertrand Clarke†**

August 27, 2004

Abstract

We use a predictive optimality criterion to search for optimal model lists. This criterion admits the usual variance bias interpretation. It also permits computations allowing empirical optimization over certain restricted classes of model lists. Our results confirm the intuition that predictive error as function of model list size typically has a V shape: The predictive error is often large for unreasonably small or simple model lists, and often large for unreasonably large or complex model lists. Predictive error usually achieves a minimum at some intermediate model list.

*Ernest Fokoué is Assistant Professor, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA (**eMail:** epf@stat.ohio-state.edu). He is also Postdoctoral Research Fellow at the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709 (epf@samsi.info).

†Bertrand Clarke is Associate Professor, Department of Statistics, University of British Columbia, Vancouver, Canada. This work was done while he was on leave visiting SAMSI and the Institute of Statistics and Decision Sciences, Duke University (**eMail:** bertrand@stat.duke.edu).

1 Introduction

Suppose we have a sequence of data from an unknown function f^*, and we want to predict the next value $f^*(\mathbf{x}^{\mathsf{new}})$. Suppose also that we have a set $\mathbb{B}$ of functions that we think will help us approximate the data generator, and have used them to obtain a model list at time t. Then, if we default to a uniform prior on models we can form a Bayes Model Average (BMA). It gives fitted values for the data and hence gives residuals. Let τ be a threshold parameter, defined formally later, controlling how readily we add more elements to the model list from which we form the BMA. Thus, the size of the model list is controlled by τ. Then, running our prediction scheme repeatedly gives the average prediction error after 50 data points as a function of τ. Figure 1 shows the results from a simulation for

$$f^*(\mathbf{x}) = 2 + 2\sin(\mathbf{x}) + 1.25\sin(2\mathbf{x}) + \sin(7\mathbf{x}) \text{ on } [-1, 1].$$

Panel (i) shows the average predictive error over 100 runs for each value of τ in $[0, 2]$ when $\mathbb{B}$ consists of the first 30 Chebyshev polynomials. A clear, strong V shape is seen. (Repeated simulations show that deviations from the V shape are due to randomness.) The minimum final average predictive error is around .7 and its maximum occurs at 2, the upper bound for τ. Panels (iii) and (iv) show the average prediction errors for $\tau = .7$ and $\tau = 2$ respectively over 10 time steps of 5 data points each. It is seen that the best τ gives a smooth drop in the average predictive error while the worst τ has poor – and more variable – performance. Panel (ii) shows the best final fit from $\tau = .7$. The worst fit from the worst τ is very poor. Panels (v) and (vi) show the worst fit from the best τ and the best fit from the worst τ.

These diagrams are consistent with a variance bias decomposition interpretation. For small τ, it is hard to add models. So, the bias is relatively large. The variance is small since there are relatively few parameters. As τ increases, the variance increases a little but the decrease in bias is bigger and the overall predictive error decreases to a minimum. Past this minimum, there are few gains from decreasing bias and greater losses from fitting too many models. Since the target function is not in the span of any finite set of polynomials the procedure keeps on adding models to the average even though they are not helpful. The predictive behavior from panels (ii) and (iii) is consistent with this interpretation: Too many models leads to instability while the right number of models permits the procedure to zero in on a good predictor. Likewise, the poor fit in panel (v) indicates bias, and panel (vi) comes from a much larger model list which will have higher variances.

We will argue below that this V shape is typical and consequently there are optimal model lists which can be identified by finding τ^{opt}. In some cases, degenerate forms of the V shape are found. In particular, there are 3 variants that arise. First, one can get analogs of panel (i) that increase from a flat region. Second, one can get analogs that decrease to a flat region. Third, one can get curves that are essentially flat (though highly variable). We regard these as degenerate forms in which either or both of the arms of the V are missing. They also permit variance bias decomposition interpretations which we defer.

Abstracting from the above, suppose we have a data generating mechanism, DGM, producing pairs $(\mathbf{x}_i, y_i)$ sequentially using an underlying true function f^*. That is, each response is of the form $Y_i = f^*(\mathbf{x}_i) + \epsilon_i$, where ϵ_i has a normal distribution with mean 0 and variance σ^2 and we assume the $\mathbf{x}_i$'s are available before the prediction is made. To approximate f^* using a set of

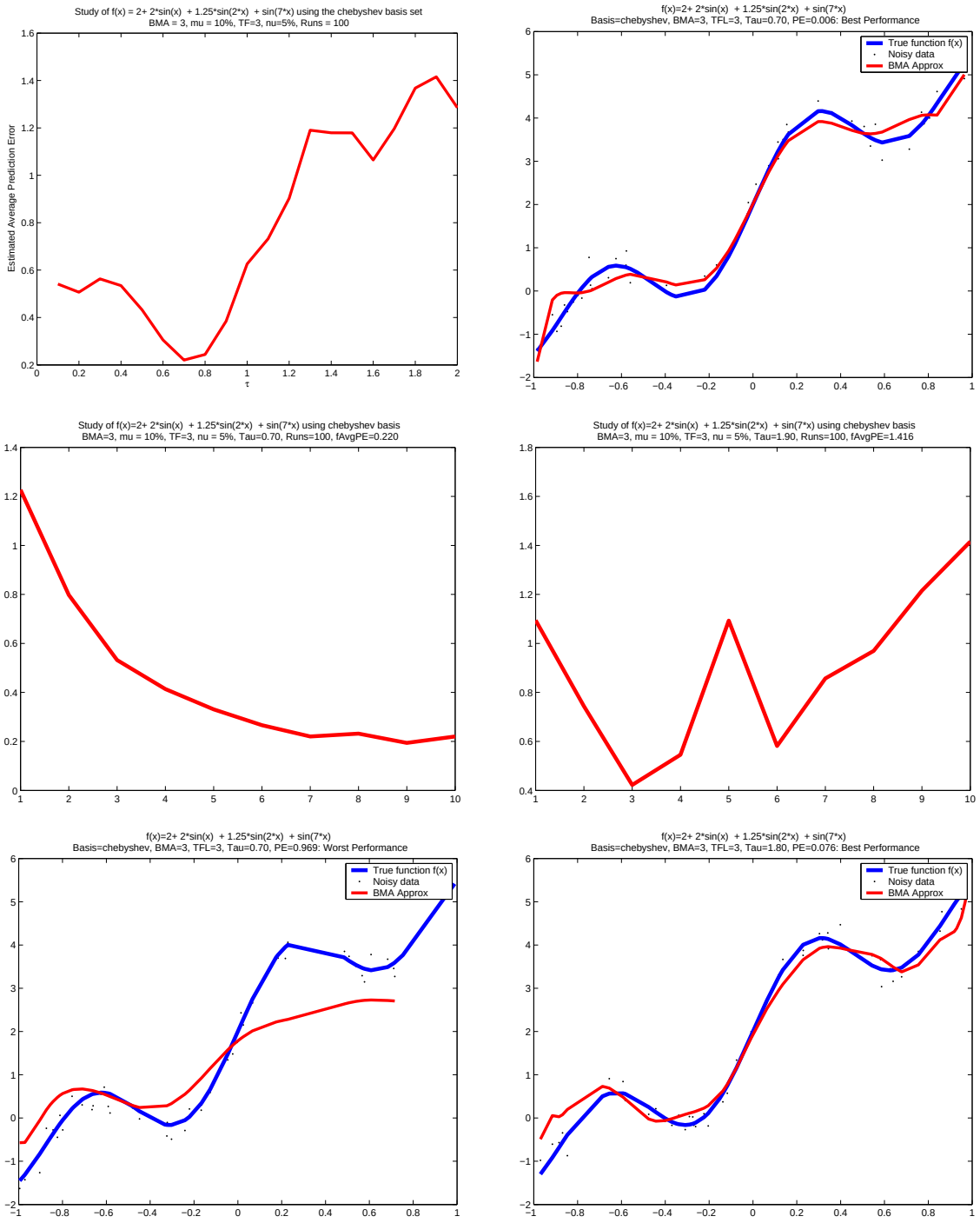


Figure 1: Hill function in the chebyshev basis with AS = 3, TF = 3

n observations, consider using models M of the form

$$Y = \beta_0 + \sum_{\gamma} \beta_\gamma B_\gamma(\mathbf{x}) + \epsilon \tag{1}$$

where each B_γ is a function of $\mathbf{x}$ from

$$\mathbb{B} = \{B_1, B_2, \cdots, B_q\}. \tag{2}$$

Typically, we choose $\mathbb{B}$ to be complete bases like Legendre or Chebyshev polynomials, or the Fourier waveforms. In our example above, we used $q = 30$ Chebyshev polynomials. In some cases, we will choose $\mathbb{B}$ to be overcomplete in the sense that a linear combination of elements in $\mathbb{B}$ may equal or approximate another element in $\mathbb{B}$. In general, $\mathbb{M}$ denotes the model space derived from $\mathbb{B}$, and $\mathcal{M}$ denotes a subset of $\mathbb{M}$, which we call a model list.

As suggested by the V formation in panel (i), our goal here is to find the optimum model list $\mathcal{M}^{\mathsf{opt}}$ that minimizes the Predictive Mean Squared Error (PMSE). i.e.,

$$\mathcal{M}^{\mathsf{opt}} \;=\; \mathsf{arg}\min_{\mathcal{M}\subset\mathbb{M}} \mathsf{PMSE}(\mathcal{M}) \tag{3}$$

where

$$\mathsf{PMSE}(\mathcal{M}) \;=\; \mathbb{E}_{\mathsf{P}_{\mathsf{true}}}\left[(y^{\mathsf{new}} - \hat{y}^{\mathsf{new}})^2\right], \tag{4}$$

in which the expectation is taken with respect to $\mathsf{P}_{\mathsf{true}}$, the density for y^{new}, and the estimate $\hat{y}^{\mathsf{new}}$ is the response predicted for $\mathbf{x}^{\mathsf{new}}$ using BMA on $\mathcal{M}$, i.e.,

$$\hat{y}^{\mathsf{new}} \equiv \mathsf{BMA}(\mathbf{x}^{\mathsf{new}}; \mathcal{M}).$$

We have chosen to use BMA because it is predictively optimal, see Barbieri and Berger (2004) for instance, and the references therein. We have defaulted to the uniform prior because we have ensured our model lists are small.

Various authors have examined the details of model list formation for BMA. For instance, Raftery et al. (1997) and Hoeting et al. (1999) used a reversible jump MCMC procedure to generate a list of models on which they could apply an Occam's window approach by thresholding the posterior model weights. Of even greater importance to our reasoning here is the foundational work of George (2000) and George and McCulloch (1993). Those authors recognized the phenomenon of dilution, and proposed priors to overcome it. Dilution priors correct for posterior skewness that may result when too many models formed from highly correlated variables split the posterior probability too finely. Despite these searches and examinations of the model space, no one seems to have formulated an optimization problem such as in (3) to formalize the concept of an optimal model list.

Here, we propose an approach to optimize over model lists using predictive optimality. A consequence of our approach is that it lets us argue that a bias-variance interpretation should guide model list selection. Our computational results are an empirical optimization under that criterion over a tractable subset of model lists.

In Section 2, we provide a short overview of BMA. In Section 3, we discuss our basis search scheme in general and its relation to other schemes. In Section 4, we present the specifics of our scheme and we describe the various scenarios in which we have generated computational results. Section 5 is a presentation of the extensive computational results we have obtained, with discussion as appropriate. Section 6 summarizes our general conclusions and indicates possible future developments. A short Appendix gives pseudo-code for our procedure.

2 An overview of BMA for normal linear models

Using predictions from only a single model will underestimate the variability of the estimate, since it ignores the fact that another model with a high posterior probability would make a different prediction. Instead, one should calculate predictions by using a weighted average over all models in the search space, where the weights are the posterior probabilities of the models (Leamer, 1978; Kass and Raftery, 1995). BMA is one way to achieve better predictions by accounting for model uncertainty. Hjort and Claeskens (2003) provide a detailed account on Frequentist Model Averaging Estimators. In general, the goal of model averaging is to improvement the fit without overfitting: Adding basis functions that do not improve the fit enough tend to give models with low posterior probability.

For models defined by equation (1), consider a sample $D^{(t)} = \{(\mathbf{x}_i, y_i), i = 1, \cdots, n_t\}$ of iid observations, and a subset $\mathcal{W}^{(t)} = \{B_{t1}, B_{t2}, \cdots, B_{tk_t}\}$ of k_t basis functions from $\mathbb{B}$. With ϵ_t following a normal distribution with mean 0 and variance σ^2 as assumed earlier, we have the normal linear model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5}$$

where $\boldsymbol{y} = (y_1, y_2, \cdots, y_{n_t})^\mathsf{T}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_{k_t})^\mathsf{T}$, $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_{n_t})^\mathsf{T}$ and

$$\mathbf{X} = \begin{bmatrix} 1 & B_{t1}(\mathbf{x}_1) & B_{t2}(\mathbf{x}_1) & \cdots & B_{tk_t}(\mathbf{x}_1) \\ 1 & B_{t1}(\mathbf{x}_2) & B_{t2}(\mathbf{x}_2) & \cdots & B_{tk_t}(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & B_{t1}(\mathbf{x}_{n_t}) & B_{t2}(\mathbf{x}_{n_t}) & \cdots & B_{tk_t}(\mathbf{x}_{n_t}) \end{bmatrix} \tag{6}$$

The likelihood is $\mathbf{y} \sim \mathcal{N}(\mathbf{V}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, or more specifically

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right]. \tag{7}$$

If we use an isotropic prior on $\boldsymbol{\beta}$ with precision parameter δ, then we can write

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \delta^{-1}\mathbf{I}_{k_t+1}). \tag{8}$$

Let y^{new} be the response variable corresponding to a new test point $\mathbf{x}^{\mathsf{new}}$. Let M_α denote the models of interest taken from a given collection of models $\mathcal{M}$. The posterior predictive distribution of y^{new} is

$$\boldsymbol{p}(y^{\mathsf{new}}|\boldsymbol{y}) = \sum_{M_\alpha \in \mathcal{M}} \boldsymbol{p}(y^{\mathsf{new}}|\boldsymbol{y}, M_\alpha)\mathbf{Pr}(M_\alpha|\boldsymbol{y}) \tag{9}$$

where $\boldsymbol{p}(y^{\mathsf{new}}|\boldsymbol{y}, M_\alpha)$ is the marginal posterior predictive density given model M_α, that is

$$\boldsymbol{p}(y^{\mathsf{new}}|\boldsymbol{y}, M_\alpha) = \int_\Theta \boldsymbol{p}(y^{\mathsf{new}}|M_\alpha, \boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{y}, M_\alpha)d\boldsymbol{\theta}, \tag{10}$$

and $\mathbf{Pr}(M_\alpha|\boldsymbol{y})$ is the posterior probability of model M_α, i.e,

$$\mathbf{Pr}(M_\alpha|\boldsymbol{y}) = \frac{\boldsymbol{p}(\boldsymbol{y}|M_\alpha)\boldsymbol{p}(M_\alpha)}{\sum_{\alpha'} \boldsymbol{p}(\boldsymbol{y}|M_{\alpha'})\boldsymbol{p}(M_{\alpha'})}, \tag{11}$$

in which

$$\boldsymbol{p}(\boldsymbol{y}|M_\alpha) = \int \boldsymbol{p}(\boldsymbol{y}|M_\alpha, \boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{\theta}|M_\alpha)d\boldsymbol{\theta} \tag{12}$$

is the marginal probability of the data. If we further assume that σ^2 and δ are known, then for our normal linear model, it is easy to show that

$$\boldsymbol{p}(\boldsymbol{y}|M_\alpha) = \mathcal{N}(\boldsymbol{y}|\mathbf{0}, \sigma^2\mathbf{I}_{n_t} + \delta\mathbf{X}_\alpha\mathbf{X}_\alpha^\mathsf{T}), \tag{13}$$

where $\mathbf{X}_\alpha$ is the design matrix for model M_α. Since we do not anticipate the availability of any information about each model in $\mathcal{M}$, we will make the usual assumption of a noninformative prior that puts equal mass on each model in $\mathcal{M}$. So, we use

$$\mathbf{Pr}(M_\alpha|\boldsymbol{y}) = \frac{\boldsymbol{p}(\boldsymbol{y}|M_\alpha)}{\sum_{\alpha'} \boldsymbol{p}(\boldsymbol{y}|M_{\alpha'})} \tag{14}$$

in place of (11).

The prediction $\hat{y}^{\mathsf{new}}$ for $\mathbf{x}^{\mathsf{new}}$ from BMA with model list $\mathcal{M}$ is therefore

$$\mathsf{BMA}(\mathbf{x}^{\mathsf{new}}, \mathcal{M}) \equiv \tilde{\beta}_0 + \sum_{\gamma=1}^{q} \tilde{\beta}_\gamma B_\gamma(\mathbf{x}^{\mathsf{new}}),$$

where

$$\tilde{\beta}_\gamma = \sum_{M_\alpha \in \mathcal{M}} \mathbf{Pr}(M_\alpha \,|\, \boldsymbol{y})\mathbb{I}(B_\gamma \in \mathcal{W}^{(t)})\hat{\beta}_\gamma$$

and

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}_\alpha^\mathsf{T}\mathbf{X}_\alpha + \sigma^2\delta\mathbf{I}_{k_t}\right]^{-1}\mathbf{X}_\alpha^\mathsf{T}\boldsymbol{y}.$$

Although assuming both σ^2 and δ are known seems unrealistic, in practice these two parameters can be estimated accurately from the full model using empirical Bayes techniques, see Barbieri and Berger (2004).

3 Basis search schemes

Our method for selecting model lists is sequential and has two main components: a basis search followed by a simple random sampling procedure over the models formed from the admitted basis elements.

3.1 Our basis search method

Our method rests on a successive reduction of distance between the target function and an the emerging approximation. Each search in our method can therefore be summarized as follows:

$$\text{select one or many } \; B_\gamma \in \mathbb{B} \; \text{ such that } \; d\left(B_\gamma, \boldsymbol{r}\right) < \tau,$$

where $d(\cdot, \cdot)$ is any suitable distance or dissimilarity measure, and τ is a threshold parameter controlling how closely one requires the candidate atom B_γ to match the residual function $\boldsymbol{r}$. In our work, we used the norm

$$d\left(B_\gamma, \boldsymbol{r}\right) \equiv \left\| \frac{B_\gamma(\mathbf{x})}{\|B_\gamma(\mathbf{x})\|} - \frac{\boldsymbol{r}(\mathbf{x})}{\|\boldsymbol{r}(\mathbf{x})\|} \right\|_p \tag{15}$$

as our distance measure where

$$\|g(\mathbf{x})\|_p \equiv \left(\int \left| g(\mathbf{x}) \right|^p d\mathbf{x} \right)^{1/p}, \tag{16}$$

is approximated for n observations

$$\|g(\mathbf{x})\|_p \approx \left[\sum_{i=1}^{n} |g(\mathbf{x}_i)|^p \right]^{1/p}. \tag{17}$$

Note that by normalizing B_γ and $\boldsymbol{r}$ before computing the norm of their difference, we have $\tau \in [0, 2]$. This boundedness will be useful when drawing conclusions from our computations.

We have chosen to use a norm because we want a magnitude that is independent of all other considerations. This is important for comparing the performance of different bases. Using an inner product directly would be sensitive to angles as well as distances, making basis to basis comparison difficult. Alternative assessments of distance, for instance a dissimilarity measure based on a kernel, may be useful in some settings, but we have not explored that here.

3.2 Other basis search methods

Another basis search method is Basis Pursuit (BP) developed by Chen et al. (2001), Chen et al. (1998) and Chen (1995). A closely related technique is the Method of Frames (MOF) proposed by Daubechies (1988). The key difference between BP and MOF is that the ℓ^1 of BP becomes an ℓ^2 in MOF. So BP usually gives sparser representations than MOF does. This parallels the difference between LASSO and ridge regression. In BP and MOF the main goal is to find a sparse functional representation of the signal of interest by solving

$$\min \|\boldsymbol{\beta}\|_p \quad \text{subject to } \mathbf{X}\boldsymbol{\beta} = \boldsymbol{y} \tag{18}$$

greedily, where $p = 1, 2$; $\mathbf{X}$ is the $n \times q$ design matrix formed using k elements of the set $\mathbb{B}$ on the n available explanatory variables. That is, Basis Pursuit seeks a representation of the signal with coefficients having minimal ℓ^1 norm.

Our basis search method differs qualitatively from both BP and MOF. First, BP and MOF are not sequential. Also, BP and MOF use the final basis selected to construct a single model, while we generate model lists in variety of ways using model averaging. In addition, they are greedy and have sparsity as their goal rather than predictive optimality. Finally, both BP and MOF require sophisticated computational methods for their search, BP even more than MOF.

3.3 Initializing and updating of the process

At each time point t, we have a subset $\mathcal{W}^{(t)} \subset \mathbb{B}$ that contains all the basis elements contributing to an ever more accurate predictive approximation. We refer to $\mathcal{W}^{(t)}$ as the working basis. We consider two ways of initializing $\mathcal{W}^{(t)}$: (a) random (b) non-random. The simplest of both is random initialization which consists in *randomly* drawing one element from $\mathbb{B}$.

$$B^{\text{init}} \equiv \text{one random draw from } \mathbb{B}. \tag{19}$$

Non-random initialization chooses from $\mathbb{B}$ the one element closest to the response variable, i.e.,

$$B^{\text{init}} \equiv \arg\min_{B_\gamma \in \mathbb{B}} d\left(\frac{B_\gamma}{\|B_\gamma\|}, \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\right). \tag{20}$$

Below, in Section 5, we show some computations using each technique. Since we do not prune the selected basis, random initialization has the disadvantage that elements may be allowed in at the start that turn out to be unrelated to the underlying function.

We consider two scenarios for updating the working basis $\mathcal{W}^{(t)}$:

☐ Adding only the best candidate within distance τ:

$$\mathcal{W}^{(t)} := \mathcal{W}^{(t)} \cup \left\{B^{\text{best}}\right\} \tag{21}$$

where

$$B^{\text{best}} \equiv \arg\min_{B_\gamma \in \mathbb{B}} \left\{d(B_\gamma, \boldsymbol{r}) : \;\; d(B_\gamma, \boldsymbol{r}) < \tau\right\}.$$

☐ Adding all the candidates within distance τ:

$$\mathcal{W}^{(t)} := \mathcal{W}^{(t)} \cup \left\{B_\gamma \in \mathbb{B} \backslash \mathcal{W}^{(t)} \;\; \text{s.t} \;\; d(B_\gamma, \boldsymbol{r}) < \tau\right\}. \tag{22}$$

For orthogonal bases these two updating techniques are equivalent. However, we prefer the form of the first because, in more general settings, the second will often admit too many similar candidate terms. The tendency of the second to over-admit terms can give collinearity. It is important to avoid this here, since our method does not include pruning.

4 Model list generation

Our method for generating candidate model lists rests on the sequential selection of basis elements. We have limited our attention to orthogonal bases because of their convenient parametric representation. They also permit the introduction of a controlled amount of overcompleteness.

4.1 Basis element selection and averaging strategies

We considered four different orthogonal basis sets, the first of which was the set of Fourier sine waveforms defined on $[0, \pi]$:

$$B_\gamma(\mathbf{x}) \equiv \sin(\gamma \mathbf{x}) \text{ with } \mathbf{x} \in [0, \pi].$$

The second was the full Fourier basis set on $[-\pi, +\pi]$:

$$B_\gamma(\mathbf{x}) \equiv \sin(\gamma \mathbf{x}) \text{ or } B_\gamma(\mathbf{x}) \equiv \cos(\gamma \mathbf{x}).$$

The third was the set of Chebyshev polynomials on $[-1, +1]$:

$$B_\gamma(\mathbf{x}) \equiv \cos(\gamma \arccos(\mathbf{x})) \text{ where } \mathbf{x} \in [-1, 1].$$

The fourth was the set of Legendre polynomials on $[-1, +1]$:

$$\begin{aligned} B_1(\mathbf{x}) &= 1 \text{ and } B_2(\mathbf{x}) = \mathbf{x}, \\ (\gamma+1)B_{\gamma+1}(\mathbf{x}) &= (2\gamma+1)\mathbf{x}B_\gamma(\mathbf{x}) - \gamma B_{\gamma-1}(\mathbf{x}) \end{aligned}$$

Once the set $\mathbb{B}$ is chosen, the sequential process initializes the working basis $\mathcal{W}^{(t)} \subset \mathbb{B}$, and subsequent iterations update $\mathcal{W}^{(t)}$ using the residuals and the search method described earlier. Our process therefore implements an automated residual analysis as a way to improve the approximation sequentially.

With $\mathcal{W}^{(t)}$, one can generate a working model space with up to $2^{k_t} - 1$ models, where $k_t = |\mathcal{W}^{(t)}|$. For small values of τ, it is very likely that k_t will also be small, and the few models in the thus-formed model space will likely all contribute well to the approximation. One might therefore be willing to retain the whole set of $2^{k_t} - 1$ models of the model list. In such cases, model list selection reduces to efficient search of the original basis set $\mathbb{B}$.

As τ gets larger, however, many of the B_γ's added are likely to contribute very little to bias correction while inflating the variance, thereby causing the prediction error to increase. When k_t gets really large as a result of a large τ, the explosive number of possible models $2^{k_t} - 1$ makes computation prohibitive and leads to inaccuracies from round-off errors. It therefore makes sense to find ways to select only a subset of $\mathbb{M}$. Many authors have used various approaches from manual screening to Ockham's razor strategies to retain only the most plausible models. However, all these methods rest on the availability of all the posterior model weights which require computationally intensive evaluations.

Considering the fact that the working basis $\mathcal{W}^{(t)}$ consists of screened basis functions that are deemed close enough to the true underlying target, we have used a random selection of a proportion μ of models from the explosive list of $2^{k_t} - 1$ models. Clearly, the first benefit here is the computational convenience. Heuristically, one could propose using

$$\mu \equiv 0.95 - \tau/2,$$

when it is positive. This means that as the model space grows larger (as measured by τ), the proportion of models drawn from it to form the model list shrinks to guarantee the model list remains computationally manageable. (We ignore problems when τ is close to 2 since such lists are formed using models with high errors, see (21) and (21).) Typically we chose values

$\mu = 0.1, 0.25$. We do not comment on this further because the value of μ seems to make little difference qualitatively; its main role was to make computations feasible.

Since we are searching for optimal model lists, we define model lists with different sizes and complexities. Consider three model averaging strategies (AS):

$$\mathsf{AS} \in \{\mathsf{small}, \mathsf{medium}, \mathsf{large}\}$$

- ☐ small: all models of size 1, 2 or 3.
- ☐ medium: all models of size $k_t/2$
- ☐ large: all models of size k_t, $k_t - 1$, or $k_t - 2$.

These model lists are ranked in order of increasing complexity, or size of their elements. It is seen that the first and third are the same size while the second is larger. The interplay between size of list and the complexity of models on it is seen in the computed results of the next section.

4.2 Term formation and overcompleteness

Clearly, sine and full Fourier are qualitatively similar as they are both trigonometric function sets. Also, as polynomials, Legendre and Chebyshev are qualitatively similar. We focused on these two classes, trigonometric and polynomial, and we explored the effect of combining basis sets from them. This allowed us to assess the gains derived from this type of overcompleteness in our context. We also considered a second type of overcompleteness, the formation of frames from complete sets (bases). We did this by extending the given basis with a few new elements formed as partial sums of its elements. Frames may contain elements that taken together are linearly dependent or in which a sum of elements may be a good approximation for another element, a sort of near overcompleteness. To form these frames, we choose three kinds of term formation (TF) strategy. They are:

1. TF=1: Use $\mathcal{W}^{(t)}$ exactly as it is, i.e, no addition (no assessment of overcompleteness)
2. TF=2: $\mathcal{W}^{(t)} := \mathcal{W}^{(t)} \cup \left\{\text{a fraction of sums of pairs from} \mathcal{W}^{(t)}\right\}$
3. TF=3: $\mathcal{W}^{(t)} := \mathcal{W}^{(t)} \cup \left\{\text{a fraction of sums of triples from} \mathcal{W}^{(t)}\right\}$

For $\mathsf{TF} \in \{2, 3\}$, the number of terms added to $\mathcal{W}^{(t)}$ can quickly became explosive. We therefore introduce an user defined extra parameter ν to control the proportion of terms added. We chose $\nu = 0.01, 0.05$. As with μ we do not comment further on ν because its value made little difference qualitatively; its main role was to make computations feasible.

4.3 The role of τ and some emerging patterns

The parameter τ, as used in the distances above, indexes how easy it is to add new basis functions that could improve the approximation. The parameter τ therefore controls both the

quality, and the size of the model lists generated from $\mathcal{W}^{(t)}$. In fact, as τ increases, our model average involves more and more models because we have not imposed any parsimony. Using τ to characterize the model list $\mathcal{M}$, we restate the goal of equation (3) as

$$\tau^{\mathsf{opt}} \quad = \quad \mathsf{arg} \min_{\tau \in [0,2]} \mathsf{PMSE}(\tau). \tag{23}$$

All the examples presented in the numerical results to follow are based on artificially generated data.

Strong advocates of parsimony would conjecture that when the target function is a finite sum of basis elements, best predictive results would be obtained asymptotically when the approximating basis is the same as the target function basis. In fact, our computation do not appear to support this, outside of two special cases noted in Secs. 5.2 and 5.3. Aside from this, the distinction $f^* \in \mathsf{span}\mathbb{B}$ versus $f^* \notin \mathsf{span}\mathbb{B}$ does not seem to affect predictive error. Clearly the second case is more typical and more realistic since, in practice, we often do not have a match between the form of f^* and the basis set used to approximate it. Moreover, even for functions that require infinitely many atoms to be approximated accurately, our computations suggest that the distinction $f^* \in \mathsf{span}\mathbb{B}$ versus $f^* \notin \mathsf{span}\mathbb{B}$ continues to be unimportant.

The two special cases occur when the model class is rich enough or the models themselves are small. With rich model classes, when $f^* \in \mathsf{span}\mathbb{B}$ some overcompleteness may be helpful. With lists of small models the bias variance tradeoff is affected by $f^* \notin Span\mathbb{B}$. In particular, when $f^* \notin \mathsf{span}\mathbb{B}$ small models in the wrong basis cannot approximate it well.

5 Details of Implementation and Numerical Results

Our procedure has 5 overall inputs. First, a target function must be given. In practice, the investigator does not know this. Here we will choose four: a hill function (in a sine basis), a valley function (in a Fourier basis), a Mexican hat function (polynomial and exponential), and a tooth function (linear plus exponential). These are in increasing difficulty. Second, an ensemble of functions must be chosen. Here we consider 4 cases. Three are bases: Fourier, Chebyshev, sine, representing trigonometric functions and polynomials. The fourth ensemble is the union of Fourier and Chebyshev. Third, we must choose one of the three TF's to decide which functions are weighted by parameters. Fourth, we choose an Averaging Strategy to decide which terms get combined. Finally, we must choose a value of $\tau \in [0, 2]$.

We generate two kinds of performance graphs. The first kind of performance graph gives the approximate final Average Prediction Error, APE, for each value of τ. The second kind of performance graph is for a fixed τ to see how the APE decreases from time step to time step to give the final APE. We did this for the values of τ that give the maximum and minimum final APE from the first performance graph.

For the hill function, we have investigated all 9 pairs (TF, AS), giving performance graphs of both kinds for each case. For the Mexican hat and tooth functions, we look only at the performance graphs of the first kind since the performance graphs of the second kind exhibit the same general shapes and phenomena as for the hill function. We look at the three AS strategies and three different bases for each. Finally, for the valley function, which involves

sines and cosines, we look only at the Fourier basis for the three AS's. The valley function needs sines and cosines unlike the hill function which only needs sines.

In the results presented below, we take outcomes in batches of 5 per time step and limit our computations to 10 time steps at most, a total of 50 data points. Recall, we have the choice of starting randomly or by initializing the procedure by selecting a certain number of ensemble elements based on the initial data that are closer than randomly chosen elements would be on average. We prefer to choose the best element from the frame to initialize the predictive process, where best means smallest norm, because this is consistent with how later elements are added. For comparison, we have used random initialization in one set of computations for the Hill function. We apply the chosen TF strategy to the initialized process, and then apply the chosen AS to the resulting terms. This gives our predictor for the data points from the next time step. The 5 residuals from the next prediction stage get used in the iterative procedure for generating the next predictor. Thus, from the 100 runs we can track how model uncertainty affects the average prediction error for the 5 data points in the last time step.

Let R be the number of repetitions, and suppose we take τ in steps of 0.1. Let $\epsilon_{r,t}$ be the prediction error at time step t during run r of the scheme. The procedure we used to generate graphs can be summarized by the following pseudo-code.

For $\tau \in [0, 2]$

  For $r = 1$ to R

    - Run algorithm Prediction
    - Extract $\epsilon_r = (\epsilon_{r,1}, \epsilon_{r,2}, \cdots, \epsilon_{r,T})$

  End

  Compute column-wise averages $\bar{\epsilon}_{.t}$ from the matrix $\epsilon = (\epsilon_{r,t})$

  Plot performance graph $\bar{\epsilon}_{.t}$ against time step t

End

Plot performance graph $\bar{\epsilon}_{.T}$ against regularizer τ

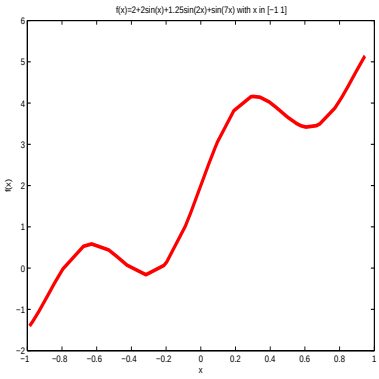
The performance graphs for fixed τ are plots of $\bar{\epsilon}_{.,t}$ versus t for the 10 time steps for which we generated predictions. Typically, they are decreasing and the speed with which they decrease indicates how difficult it is to approximate the target function with the frame used. The performance graphs that plot $\bar{\epsilon}_{.,T}$ as a function of τ often have a mode and the minimizing value of τ indicates the optimal reduction of bias.

In all our examples below, we have used a noise variance $\sigma^2 = (0.2)^2$. We chose this value because it ensured a good trade-off between identifying the target function and retaining enough randomness. Thus, in the limit of many runs of large length the lower bound for the average prediction error is $\sigma^2 = .04$. In each case we have chosen a function to approximate, a basis in which to approximate it, an averaging strategy, a term formation strategy and a value of τ.

5.1 Example 1: The Hill function

As a first example, let f^* be the Hill function used in the introduction and given by

$$f^*(\mathbf{x}) = 2 + 2\sin(\mathbf{x}) + 1.25\sin(2\mathbf{x}) + \sin(7\mathbf{x}) \text{ on } [-1, 1].$$



First, we consider the sine basis. For each of the 9 (TF, AS) pairs, we give the first performance graphs and then the performance graphs of the second kind for the best and worst τ's. Then we give the corresponding graphs using the Chebyshev basis.

5.1.1 The Hill function sine basis

In Fig. 2, the rows correspond to TF=1,2,3 for AS=1. The first column shows the performance graph of the first kind; the two following columns show the performance graphs of the second kind for the τ achieving the least and greatest final APE, respectively. The first column shows that the graphs do not depend on the TF strategy. This suggests that term formation does not affect the variance bias tradeoff in this predictive setting. It is seen that the APE over time decreases for the best τ and increases for the worst τ. The performance graphs of the second kind for the worst τ also appear more erratic. This may indicate that bad τ's permit good models and bad models indiscriminately, leading to high variability in prediction as a consequence of unregulated bias.

It is seen that the performance curves of the first kind decrease from a peak to a minimizing value and stay there as τ increases. This is a degenerate V in which the right arm increase does not happen because the models are small. Such models do not tend to overfit.

We comment that the increase at the beginning of some of the performance graphs of the first kind, here and in Figures 3 and 4, is the consequence of random initialization. That is, when we used a random initialization, we typically got a short-lived, rapid increase in APE as the procedure locked onto models that gave improved predictions. In Figures 5, 6, and 7, where we used non-random initialization, this increase does not appear. In Fig. 3, the rows correspond to TF=1,2,3, as before, but for AS=2. The first column shows the performance graph of the first kind; the two following columns show the performance graphs of the second kind for the τ achieving the least and greatest final APE, respectively.

The first column shows a dependence on TF. As TF increases, the strength of the V-formation increases. This is the only case we found where the value of TF affected the results. We suspect

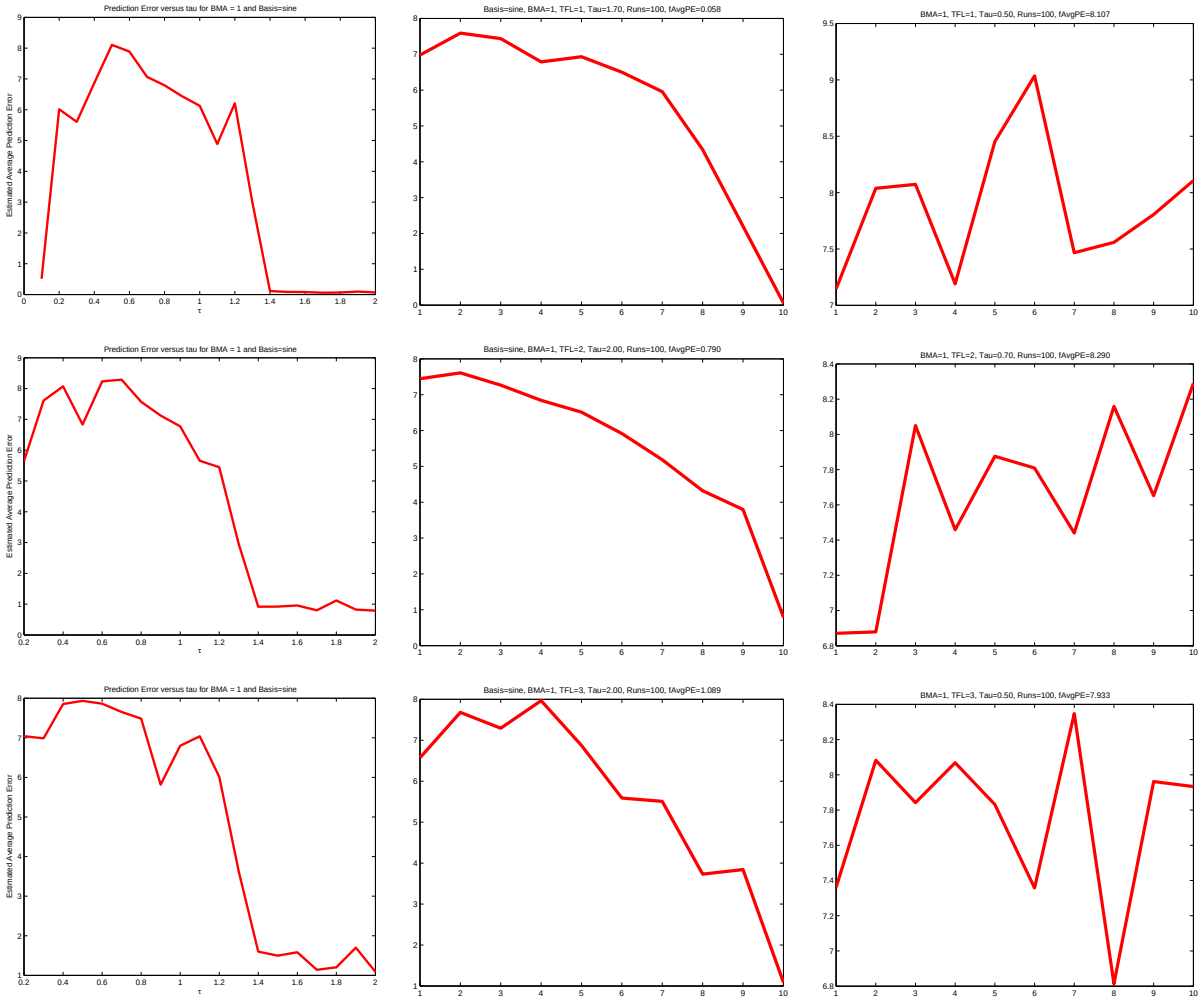


Figure 2: Hill function in the sine basis with AS = 1

this is not random variability because the model class for AS=2 has mid-sized models: They have $k/2$ terms midway between the fewest terms, $k = 1, 2, 3$ and the maximal numbers of terms $k, k-1, k-2$. Since there are more mid-sized models than large or small ones, and the basis is the same as that of the target function, it may be that term formation permits a faster search of the models, as expected from using overcompleteness. Indeed, it is seen that the APE over time decreases for the worst τ's, as well as the best ones. In Fig. 4, the rows correspond

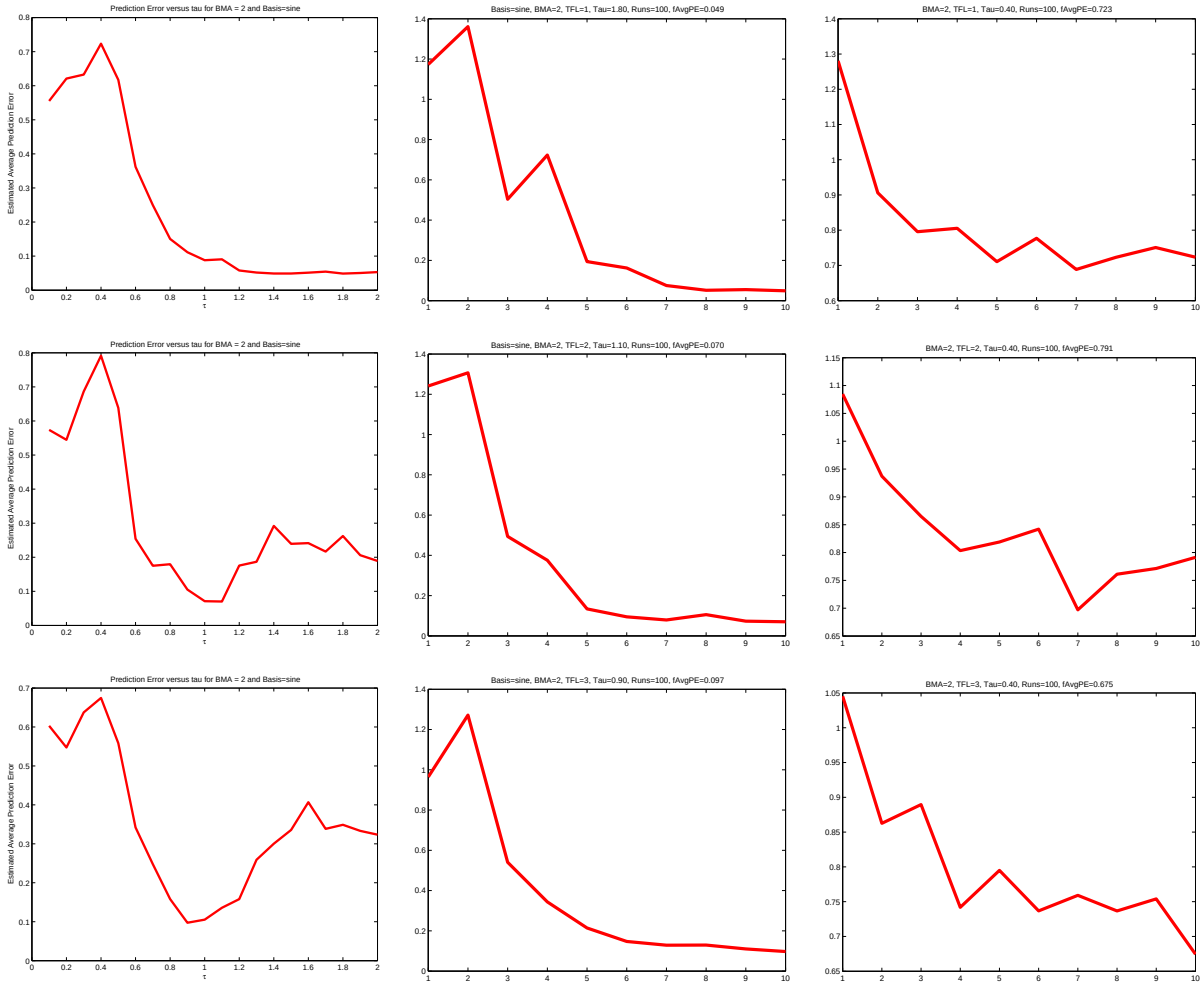


Figure 3: Hill function in the sine basis with AS = 2

to TF=1,2,3, as before, but for AS=3. The first column shows the performance graph of the first kind; the two following columns show the performance graphs of the second kind for the τ achieving the least and greatest final APE, respectively.

Like Fig. 2, the first column shows no dependence on TF and the performance curves of the first kind decrease from a peak to a minimizing value and stay there as τ increases. As before, the right arm of this degenerate V does not increase because the models are small. Such models do not tend to overfit. Averaging over models that do not overfit will not give errors typical of overfitting. Also as before, we see a smooth decrease in the column for the best τ's and high variability in the column from the worst τ's.

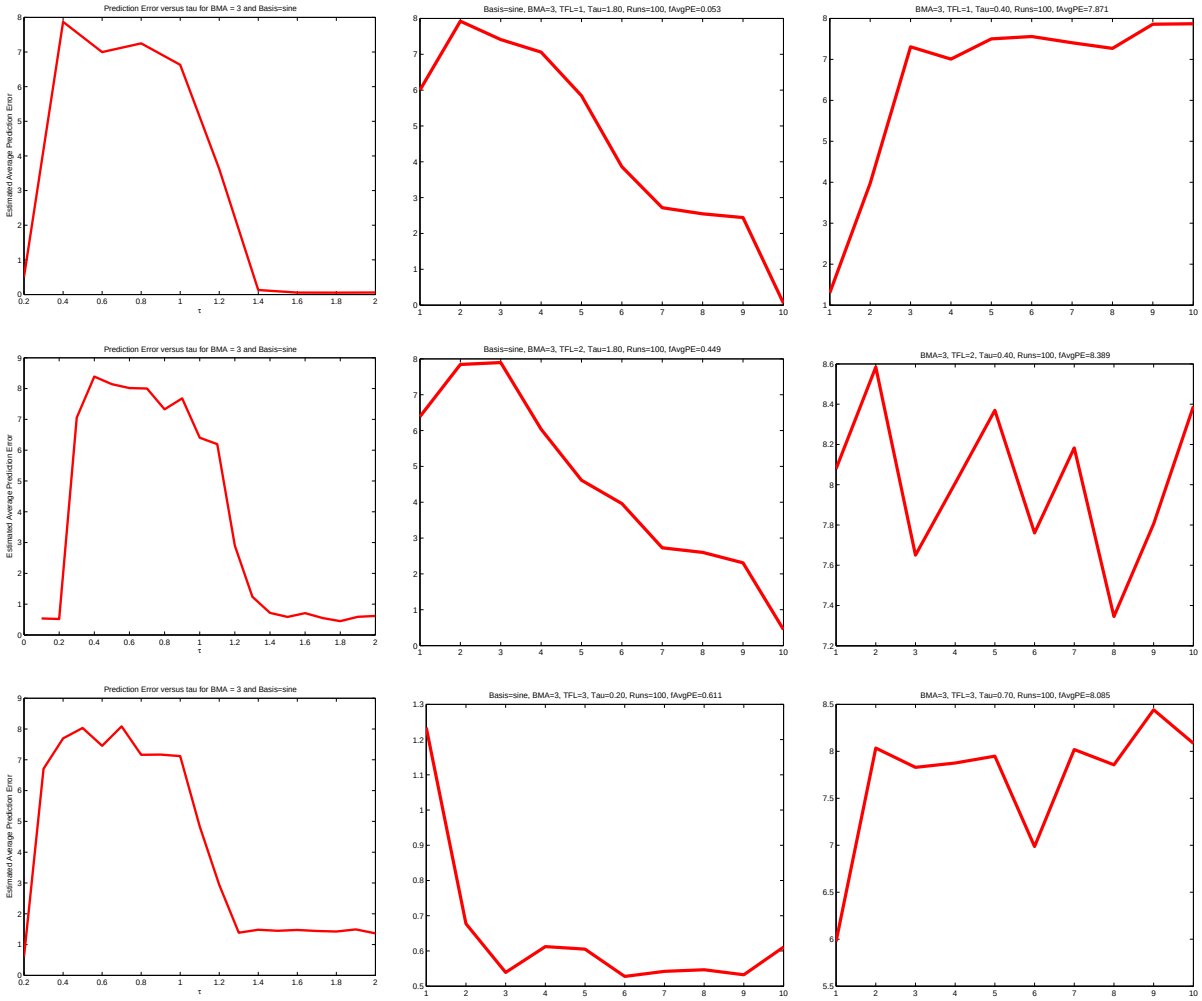


Figure 4: Hill function in the sine basis with AS = 3

5.1.2 The Hill function in chebyshev basis

Here, we have redone the computations from the previous subsection using the Chebyshev basis in place of the sine basis. We have also used non-random initialization on the working basis.

In Fig. 5, TF does not appear to make a difference. The common appearance of the performance graphs of the first kind is increasing. This means that small models in the wrong basis cannot approximate the target function well. As before, the best and worst τ's show a decrease and increase in APE, respectively. The smoothness in the worst τ case means that good models are essentially never found, whereas earlier good models could occasionally be found. This is a degenerate V in which the left arm is flat because the bias is not high enough for small τ's. In

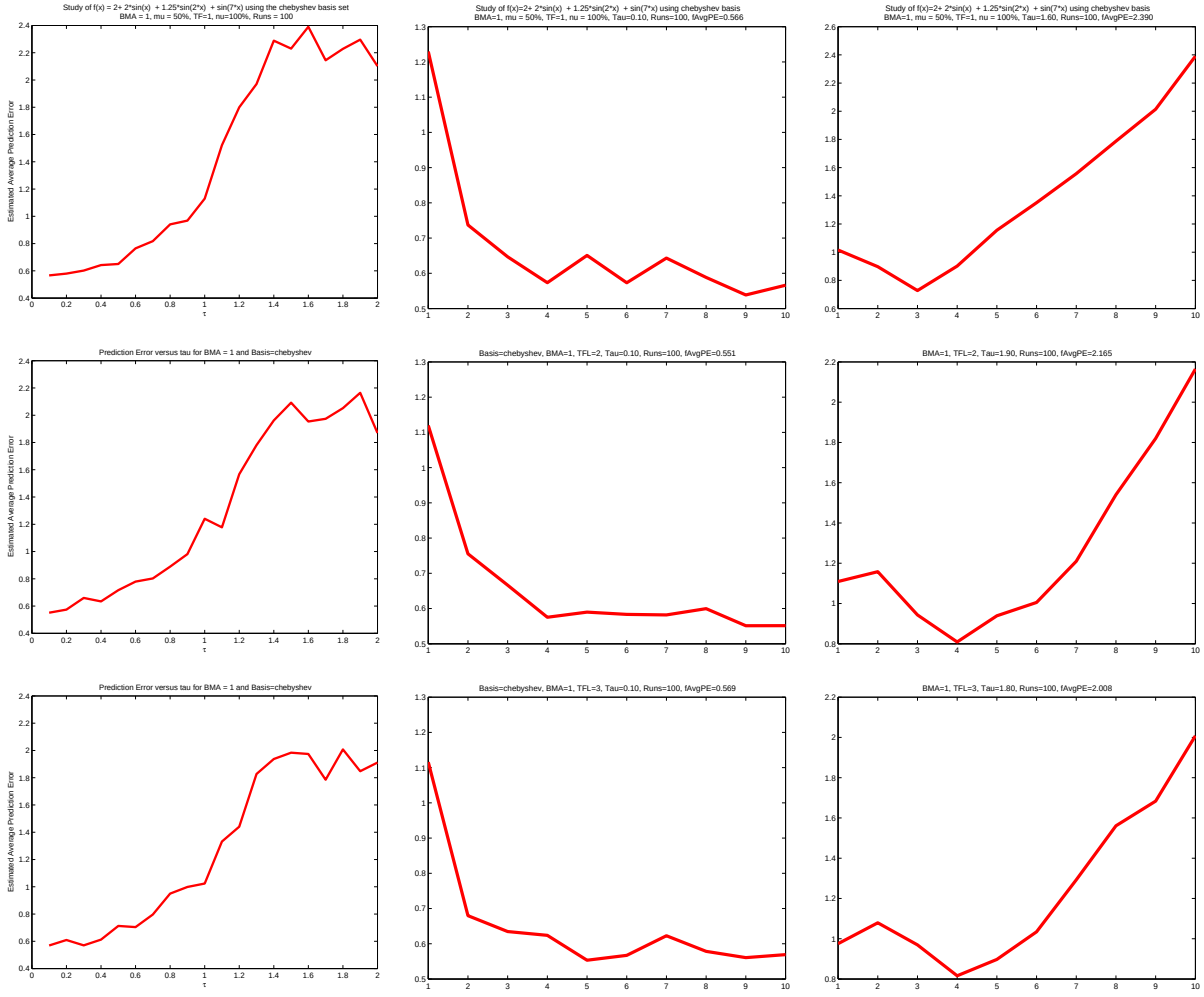


Figure 5: Hill function in the chebyshev basis with AS = 1

Fig. 6, TF does not appear to make a difference. The common appearance of the performance graphs of the first kind is increasing to a level around which there is high variability. This means that the models AS=2 uses, when the basis ins wrong, give good approximations sometimes and bad approximations others. As before, the best and worst τ's show a decrease and increase in APE, respectively. The third column indicates that we are getting problems with overfit for the worst value of τ. We do not attribute this to TF. Rather we regard the V-formation in the

worst τ performance graph for TF=2 as the result of model overfitting. This contrasts with the *model list overfitting* that high values on the right arms of V-formations in the first performance graphs indicate. Again, this is a degenerate V lacking a left arm. In Fig. 7, TF does not appear

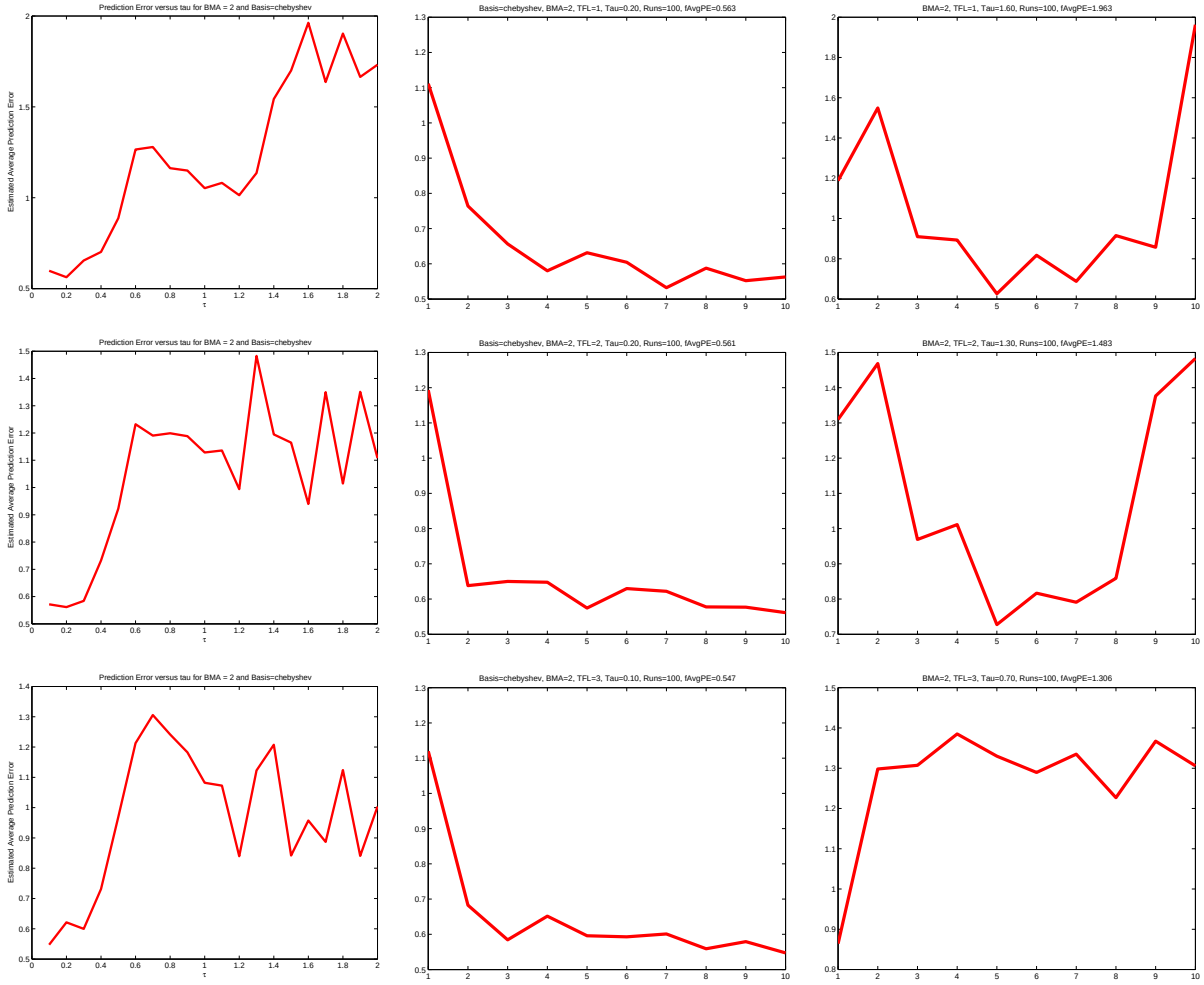


Figure 6: Hill function in the chebyshev basis with AS = 2

to make a difference. The common appearance of the performance graphs of the first kind is a V-formation. The models here for AS=3 are large, though not as numerous as for AS=2. So, each model on the list is big enough to provide good approximation. So, as more models are used APE decreases until problems with overfit emerge. Overfit from individual models is leading to overfitting of the model list. This pattern is the defining feature of good model list selection. The performance graphs of the second kind admit interpretations as before.

5.2 The Mexican Hat function

The function

$$f^*(\mathbf{x}) = (1 - \mathbf{x}^2)\exp(-0.5\mathbf{x}^2) \qquad \mathbf{x} \in [-2\pi, 2\pi]$$

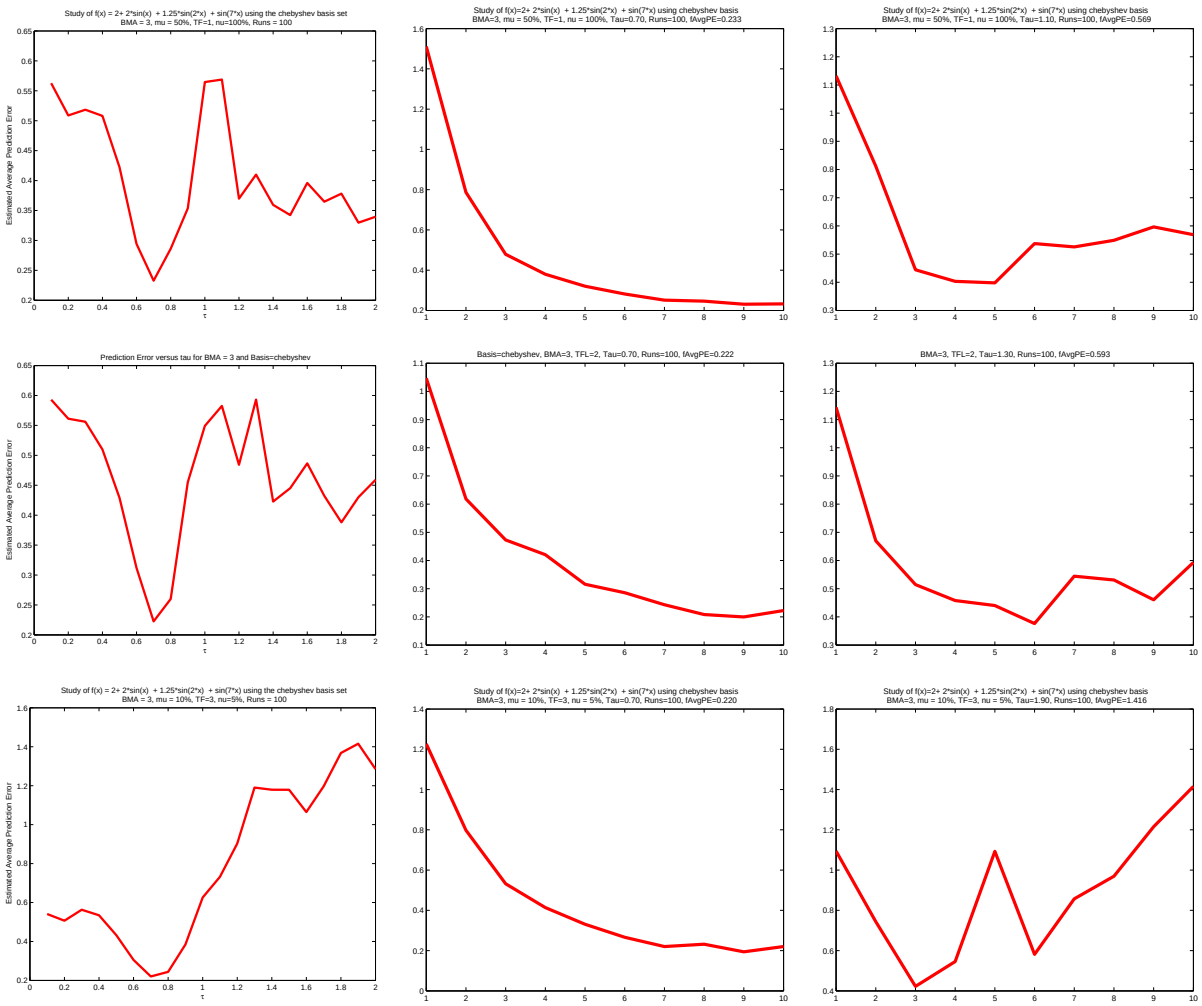
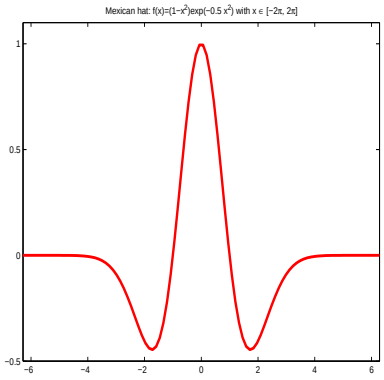
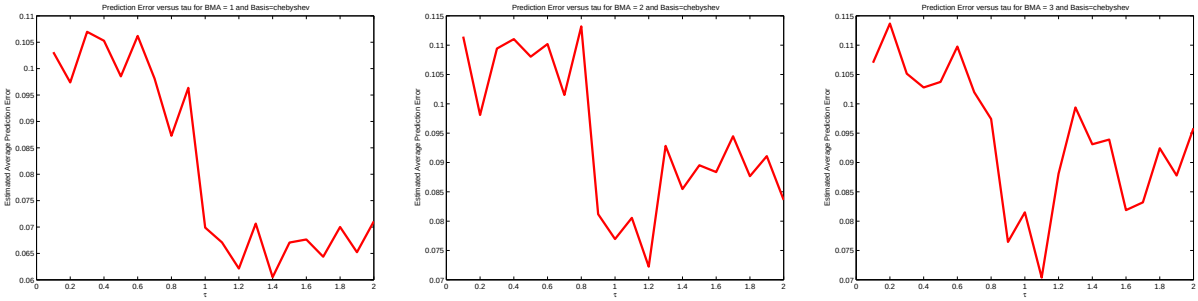


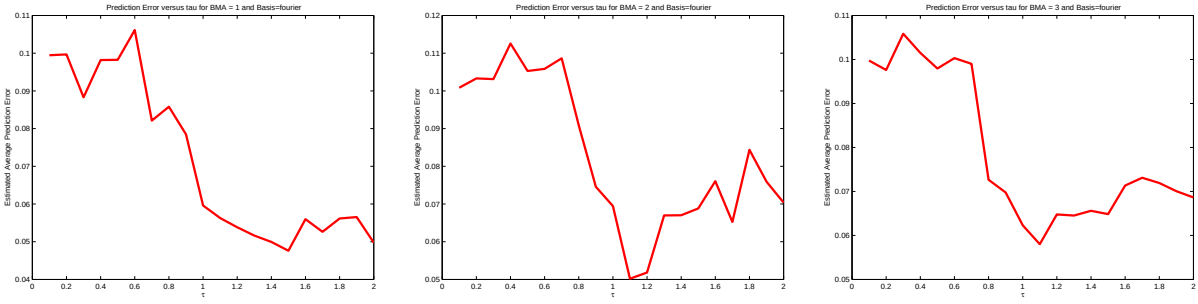
Figure 7: Hill function in the chebyshev basis with AS = 3

shown below, is not expressible with finitely many elements from the bases we have considered here. However it does have a point of symmetry that may make it less complex than the tooth function we use in the next section.

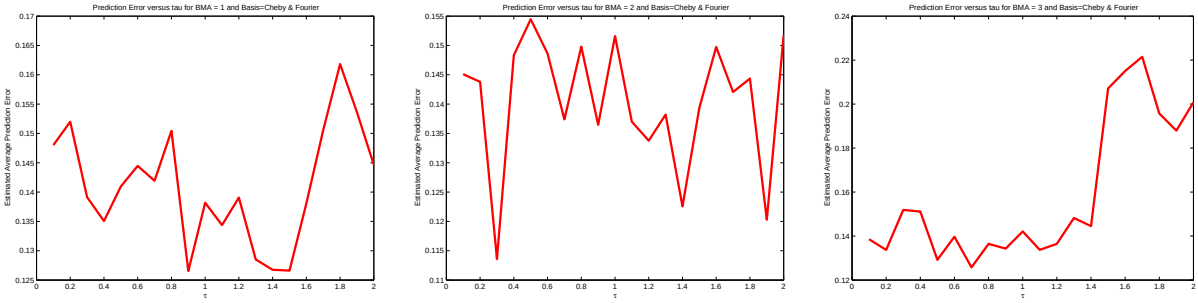




(a) AS $\in \{1, 2, 3\}$ in Chebyshev



(b) AS $\in \{1, 2, 3\}$ in Fourier



(c) AS $\in \{1, 2, 3\}$ in Chebyshev Fourier combined

Figure 8: Mexican hat function in the chebyshev and fourier bases

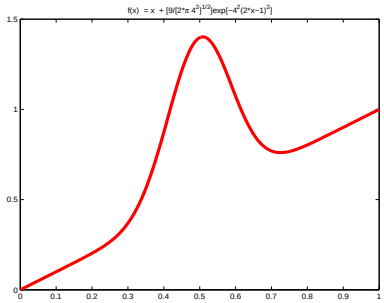
The results for the Mexican hat are shown in Figure (8). Rows correspond to the bases Chebyshev, Fourier and Combined Chebyshev and Fourier; columns correspond to $\mathsf{AS} \in \{1, 2, 3\}$. All are first performance curves of APE over $\tau \in [0, 2]$. It is seen that the first performance graphs exhibit, to varying degrees, the expected V-formation. In the top 2 rows, as AS increases, the strength of the V-formation increases as expected. When the bases are combined, small and large models evidence a weak V. Mid size models do not appear able to discriminate over model richness as summarized by τ. It is seen that the trend line from the graph is flat, in spite of the highly variable appearance. We suspect the lack of discrimination and concomitant increased variability reflects the increased richness of the models from the combination of bases.

5.3 Tooth Function

The function

$$f^*(\mathbf{x}) = \mathbf{x} + \frac{9}{4\sqrt{2\pi}} \exp\left[-4^2(2\mathbf{x} - 1)^2\right] \qquad \mathbf{x} \in [0, 1]$$

shown below, is not expressible with finitely many elements from the bases we have considered here. In addition, it is asymmetric. However, it is localized.



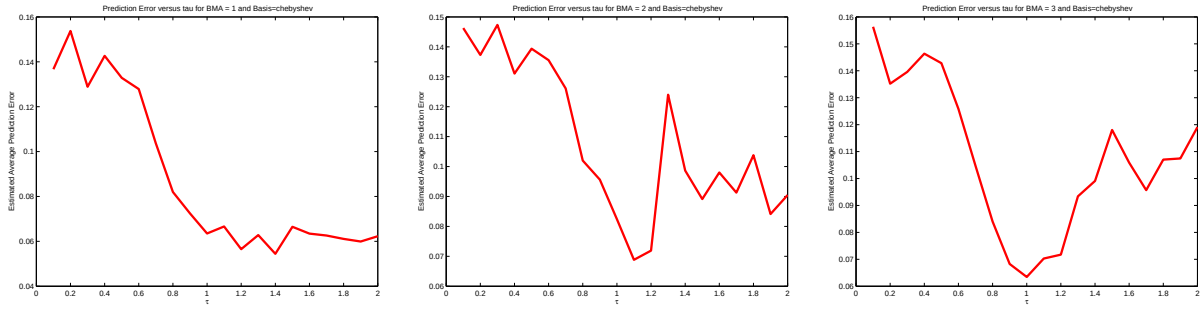
The results for the tooth are shown in Figure (9). Again, rows correspond to the bases Chebyshev, Fourier and Combined Chebyshev and Fourier; columns correspond to $\mathsf{AS} \in \{1, 2, 3\}$. All are first performance curves of APE over $\tau \in [0, 2]$. It is seen that the first performance graphs exhibit, to varying degrees, the expected V-formation. For each basis, the V-formation strengthens as AS increases. For all bases, the $\mathsf{AS}{=}1$ column shows a degenerate V that does not increase as τ gets close to 2. This may indicate that for highly localized functions such as the tooth, small models fit so poorly that even when it is easy to add them they are added so rarely that overfit does not occur.

5.4 Valley Function

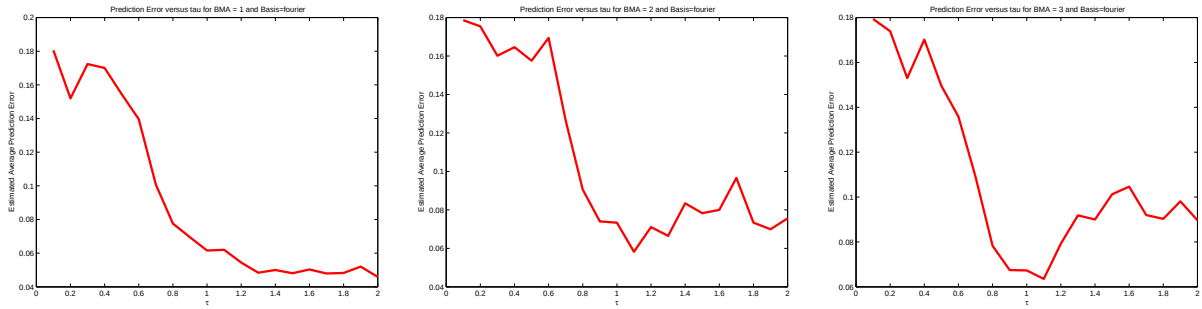
The function

$$f^*(\mathbf{x}) = 2 + \sin(\mathbf{x}) + 0.5\cos(\mathbf{x}) \quad \mathbf{x} \in [-\pi, \pi]$$

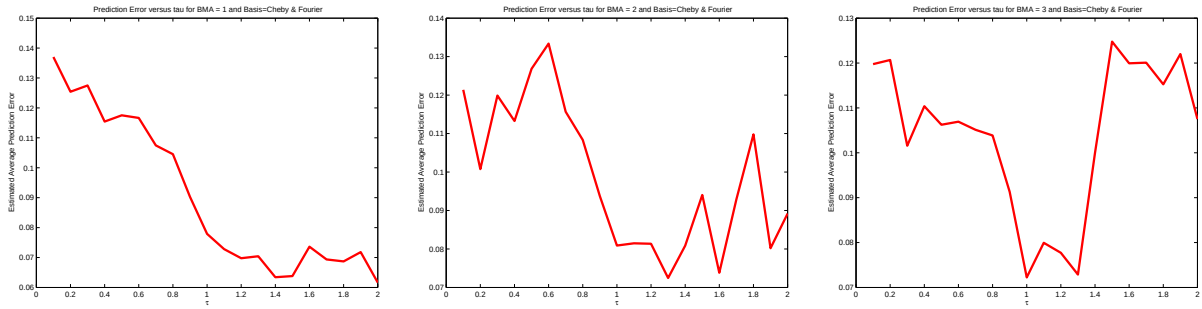
shown below, is expressible with finitely many terms from the Fourier basis. We have included it for contrast with the hill function expressible in the sine basis.



(a) AS $\in \{1, 2, 3\}$ in Chebyshev



(b) AS $\in \{1, 2, 3\}$ in Fourier



(c) AS $\in \{1, 2, 3\}$ in Chebyshev Fourier combined

Figure 9: Tooth function in the chebyshev and fourier bases

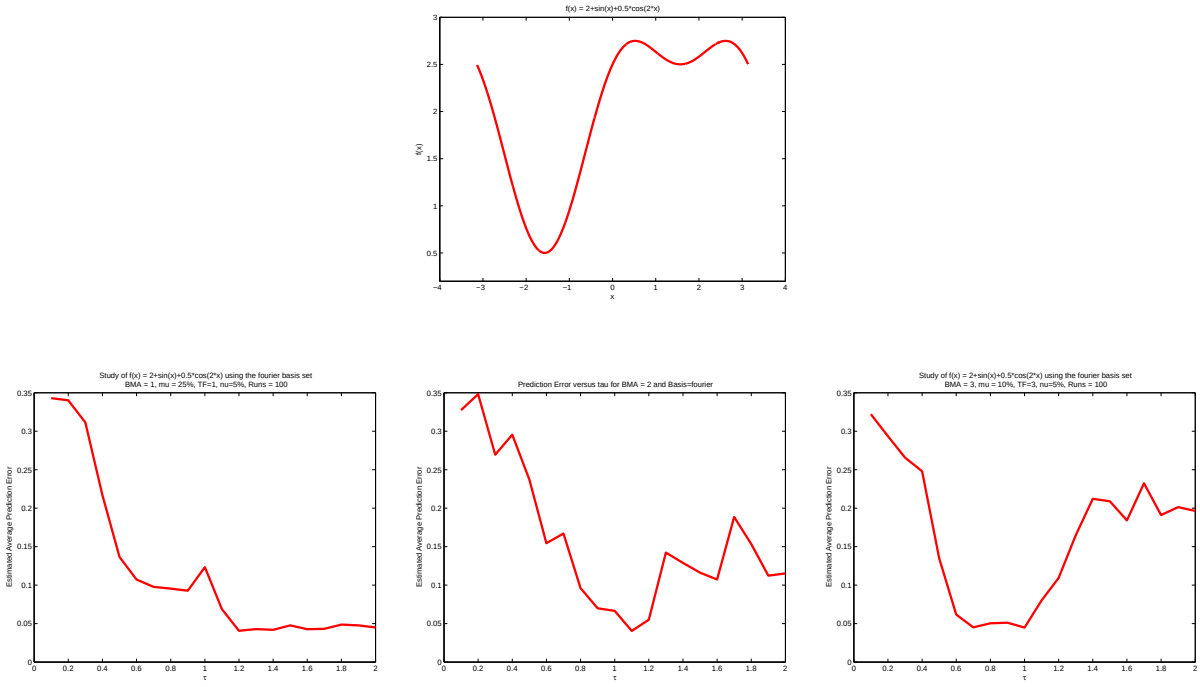


Figure 10: First performance curves for the Valley function with $\mathsf{AS} \in \{1, 2, 3\}$, in Fourier

The results for the valley function in the Fourier basis are shown in Figure (10). From left to right, we have used $\mathsf{AS} \in \{1, 2, 3\}$. Again, the V-formation strengthens as AS increases. Also as before, the case $\mathsf{AS}{=}1$ shows a degenerate V that does not increase as τ gets close to 2. As with the tooth function, this may indicate that for localized functions, small models fit so poorly that even when it is easy to add them they are added so rarely that overfit does not occur.

6 Conclusion and Discussion

Every performance graph of the first kind we have shown corresponds to a V formation, or a degenerate V formation. Each of the V-formations has a unique variance bias tradeoff interpretation on the level of model lists. For instance, a decreasing pattern down to a minimal value at which the curve is constant only occurs for small model lists, AS=1 and AS=3 for instance are smaller lists than AS=2. (The models in AS=3 are larger than in AS=1,2 but fewer in number.) Thus, all decreasing patterns occur with them, except for hill in its own basis with TF=1 which we explain by overcompleteness. Hill in its own basis with AS=3 is a bit of an exception in the sense that when we used other functions we found a proper V formation as expected. This partial exception may reflect the sine basis is parsimonious for hill and so can stabilize.

An increasing pattern often occurs when the basis used is wrong. It is exacerbated when the models are relatively small, AS=1, and sometimes when AS=2. Both situations make it difficult to include as many terms as necessary for good function approximation.

A complete V formation tends to appear weakly with AS=2, and strongly with AS=3. The exceptions are hill in its basis which we have already discussed and the Mexican hat with AS=3 when Fourier and Chebyshev are combined. This latter case may be a very weak V, but we suggest that combining the bases reduces reduces bias so quickly that it seems constant initially, until increasing variance causes the rise in APE.

The role of term formation was surprisingly small, only playing a role when the basis was right and the models complex.

The size of τ controls not just the size of the model list, but also how well the included models fit. That is, τ plays a role in model list selection as well as model selection for the lists. This dual role of τ may be at the heart of the exceptions noted above; the quantity that τ controls, $d(B_\gamma, \boldsymbol{r})$, is a function of $\boldsymbol{r}$ which comes from BMA on a model list which τ has helped choose sequentially.

As a generality, approximation methods do better when there is a mechanism for removing basis elements or models that prove to be of little or no use. We have not included pruning here because our goal has been to search for optimal model lists rather than optimal models. We have not pruned models because we are selecting them not for their individual predictive ability, but rather to improve the average over the list. Not pruning out a basis element from one model may help it combine with another model to give better prediction. Alternatively, a basis element that is bad for one model may be very good for another.

For exactitude, we comment that we have not really used BMA. We have used posterior weights formed from a uniform prior at each time step. That is, from time step to time step we have not updated the posterior according to Bayes rule. In fact, model list reselection is akin to using a data dependent prior. In effect, we restart the decision problem at each time step with a uniform prior on a new, reselected, model list and use BMA, the optimal Bayes action for it. Our use of the uniform prior is, at best, an approximation to good prior selection. However, prior selection matters most in the 'tails' of the model list, and our model lists are small enough that tail behavior will not overwhelm the validity of our conclusions.

A natural extension of our work here would be to use full models directly in place of basis elements. If we are given a collection of models $\mathbb{M} \equiv \{M_1, M_2, \cdots, \}$ we can use M_γ in place of B_γ so that $d(M_\gamma, \boldsymbol{r}) < \tau$ becomes our inclusion criterion. If we have $\mathbb{M}$, it represents extra information which we would want to use; it would corresponding to knowing the right basis. Our goal has been to automate human understanding with weak assumptions. A consequence is that we get higher model list variability because of our basis element selection than we would get had we automated the generation of whole models from a physically justified mode list rather than terms from a proposed basis.

It is important to recognize that model averaging with one-term models that happen to be basis elements is not the same as the basis expansion for the function with those elements. A basis expansion finds the unique coefficients roughly corresponding to the projection of the function onto the space spanned by those elements. By contrast, a model average finds the convex combination of the one-term models closest to the function. The two will only be equivalent when the projection onto the span of the elements is a convex combination of the elements.

A limiting factor on the richness of model lists is the phenomenon of dilution, see George (2000). The problem of dilution can arise when there may be a large number of models that provide roughly equally good fits for the data. This is particularly a problem with correlated

explanatory variables. When BMA is used in these cases, the posterior weights may be asymmetric and so lead to a skewed average because there are regions of high redundancy in model space. A skewed average will be far from its mean and so not be representative of its average. In the linear models context, George (2000) has proposed assigning uniform prior probabilities to neighborhoods defined by Voronoi tesselation of the model space. We have sought to avoid dilution by limiting the number and complexity of models we consider. Also, we have used a transformed univariate input space rather than confronting multiple regression directly.

Another limiting factor on model list size is the phenomenon of vague convergence to zero. The problem is that when the model list gets too large relative to the data, all the posterior weights tend to get small. In these cases, it is possible that the posterior probability is spread so thinly over a large region that predictions are driven mostly by the prior rather than by the data. In extreme cases, as the prior spreads over larger and larger model lists the posterior distribution may converge pointwise to zero. We found this problem computationally in many cases.

Taken together the computational results presented here are broadly consistent with a generalized bias variance formula. Although we cannot identify this explicitly, we posit there is a tradeoff among the various components of variance and bias across modelling levels.

In our work here there are at least 5 identifiable sources of variability. They are

- ☐ Parameter uncertainty: addressed through Bayesian Parameter Estimation.
- ☐ Model uncertainty: addressed through Bayesian Model Selection and Bayesian Model
- ☐ Model List uncertainty: addressed through sequential model list selection
- ☐ Basis Uncertainty: addressed through merging.
- ☐ Model Space uncertainty: addressed through choice of basis, TF and averaging strategy.

Each contributes a variance and a bias to the overall prediction. So, we can imagine a list of 5 variance components and 5 bias components that sum to express a total variability to be associated with the prediction scheme. We suggest that if a suitable expression can be obtained, it could be optimized, in parallel to existing optimizations of mean squared error, to give optimal model lists such as we have sought here.

Acknowledgments:
The authors gratefully acknowledge the SAMSI data mining participants, especially Prem Goel, for contributing their insights.

References

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *Annals of Statistics 32*.

Chen, S. S. (1995). *Basis Pursuit*. Ph. D. thesis, Department of Statistics, Stanford University, (http://www-stat.stanford.edu/schen).

Chen, S. S., D. L. Donoho, and M. A. Saunders (1998). Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing 20*, 33–61.

Chen, S. S., D. L. Donoho, and M. A. Saunders (2001). Atomic Decomposition by Basis Pursuit. *SIAM Review 43*, 129–159.

Daubechies, I. (1988). Time-Frequency Localization Operators: A Geometric Phase Space Approach. *IEEE Transactions on Information Theory 34*, 605–612.

George, E. I. (2000). The Variable Selection Problem. *Journal of the American Statistical Association 95*, 1304–1308.

George, E. I. and R. E. McCulloch (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association 88*, 881–889.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Association 98*, 879–899.

Hoeting, J. A., D. Madigan, A. Raftery, A. E., and C. T. Volinsky (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science 14*, 382–401.

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

A Detailed description of the procedure

Prediction : **Constructing the optimal basis set**

- Get the first batch of data and call it: $\{(\mathbf{x}_i, y_i), i = 1, \cdots, m\}$
- Choose term formation index TF and model averaging index AS
- Choose an accuracy level ε and threshold τ.
- Initialize $t := 0$ and Initialize $\mathcal{W}^{(0)}$ with two randomly picked elements of $\mathbb{B}$.
- Repeat
  - Get next batch $t := t + 1$
  - Form frame $\mathcal{B}^{(t)} :=$ TermFormation$(\mathcal{W}^{(t)}, \mathsf{TF})$
  - Construct the new approximation $\mathsf{BMA}^{(t)}(\cdot)$ from $\mathcal{B}^{(t)}$ and AS.
  - Estimate the response for current data:

    $\hat{y}_i = \mathsf{BMA}^{(t)}(\mathbf{x}_i) \quad$ for $\; i = 1, \cdots, mt$
  - Compute the first order residuals:

    $r_i = y_i - \hat{y}_i = y_i - \mathsf{BMA}^{(t)}(\mathbf{x}_i) \quad$ for $\; i = 1, \cdots, mt$
  - Search $\mathbb{B} \backslash \mathcal{W}^{(t)}$ for all the B_j that capture the structure left unexplained by $\mathsf{BMA}^{(t)}$.
    These are the B_j's that are closest to $\boldsymbol{r}$ according to a distance measure and a regularizing threshold τ.

    for $j = 1$ to $|\mathbb{B} \backslash \mathcal{W}^{(t)}|$
    Compute $\boldsymbol{r} := (r_1, r_2, \cdots, r_{mt})^{\mathsf{T}}$
    Compute $\mathbf{B}_j := (B_j(\mathbf{x}_1), B_j(\mathbf{x}_2), \cdots, B_j(\mathbf{x}_{mt}))^{\mathsf{T}}$
    Compute $\rho_j := d(\mathbf{B}_j, \boldsymbol{r})$
    if $\rho_j \leq \tau$ then $\mathcal{W}^{(t)} := \mathcal{W}^{(t)} \cup \{B_j\}$
    end
  - Form the new frame $\mathcal{B}^{(t)} :=$ TermFormation$(\mathcal{W}^{(t)}, \mathsf{TF})$
  - Construct the new approximation $\mathsf{BMA}^{(t)}(\cdot)$ from $\mathcal{B}^{(t)}$
  - Assess the predictive accuracy yielded by $\mathcal{B}^{(t)}$
    . Get the next batch of data as your test set: $\{(\mathbf{x}_i^{(t+1)}, y_i^{(t+1)}), i = 1, \cdots, m\}$
    . Compute the prediction error: $\hat{R}(f, \mathcal{B}^{(t)}) = \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} |y_i^{(t+1)} - \mathsf{BMA}^{(t)}(\mathbf{x}_i^{(t+1)})|^2$
- Until $\quad (t = T - 1) \quad$ or $\quad |\hat{R}(f, \mathcal{B}^{(t-1)}) - \hat{R}(f, \mathcal{B}^{(t)})| < \varepsilon$

Note: *The new B_j's provide the main ingredient for bias correction since they help search beyond the space covered by the current model* $\mathsf{BMA}^{(t)}$.