7-2011

Stable Radial Basis Function Selection via Mixture Modelling of the Sample Path

Ernest Fokoue
Rochester Institute of Technology

Recommended Citation
Fokoue, E. (2011). Stable radial basis function selection via mixture modelling of the sample path. Journal of Data Science 9(3), 359-372.

Stable Radial Basis Function Selection via Mixture Modelling of the Sample Path

Ernest Fokoué*

Center for Quality and Applied Statistics
Rochester Institute of Technology
98 Lomb Memorial Drive, Rochester, NY 14623
e-mail: ernest.fokoue@rit.edu

Abstract: We consider a fully Bayesian treatment of radial basis function regression, and propose a solution to the the instability of basis selection. Indeed, when bases are selected solely according to the magnitude of their posterior inclusion probabilities, it is often the case that many bases in the same neighborhood end up getting selected leading to redundancy and ultimately inaccuracy of the representation. In this paper, we propose a straightforward solution to the problem based on post-processing the sample path yielded by the model space search technique. Specifically, we perform an a posteriori model-based clustering of the sample path via a mixture of Gaussians, and then select the points closer to the means of the Gaussians. Our solution is found to be more stable and yields a better performance on simulated and real tasks.

AMS 2000 subject classifications: Primary 60K35; secondary 60K35
.

Keywords and phrases: High-dimensional Function Approximation, Radial Basis Functions, Kernels, Optimal Prediction, Bayesian model selection, Mixture Modelling.

1. Introduction

We are given a training set $\mathcal{D} = \{(\mathbf{x}_i, \mathrm{y}_i), i = 1, \cdots, n : \ \mathbf{x}_i \in \mathcal{X} \subset I\!R^p, \mathrm{y}_i \in I\!R\}$, where the y_i's are realizations of $Y_i = f^*(\mathbf{x}_i) + \epsilon_i$, with the ϵ_i's representing the noise terms, herein assumed to be independently normally distributed with mean 0 and variance σ^2. Our goal is to use the information contained in the data $\mathcal{D}$ to build an estimator $\hat{f}_n$ of the true unknown function f^* that achieves the smallest mean squared error. It turns out that, without some extra knowledge or at least assumptions about some properties of f^* like its *smoothness*, finding a decent estimator is a task that belongs in the category of ill-posed problems. Typically, one assumes that f^* belongs to some normed function space, say $\mathcal{H}$, in which a global property of f^* like its smoothness is defined. We therefore need to find

$$\arg\min_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{n} (\mathrm{y}_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} \tag{1.1}$$

where $\lambda > 0$ is known as the smoothing parameter or regularization parameter, chosen to achieve a trade-off between approximation and interpolation. From a practical and/or computational perspective, it is natural to wonder what kind of function spaces can be used, and how to compute the desired estimator of f^* in that space. In this paper, we assume that $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), meaning that $\mathcal{H}$ is a Hilbert space of functions $f : \mathcal{X} \to I\!R$ equipped with a unique kernel $K : \mathcal{X} \times \mathcal{X} \to I\!R$, satisfying

*Ernest Fokoué is Assistant Professor of Statistics at Rochester Institute of Technology, USA.

1. $K(\cdot, \mathbf{x}) \in \mathcal{H}$ for all $\mathbf{x} \in \mathcal{X}$,
2. $f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$ for all $\mathbf{x} \in \mathcal{X}$ and all $f \in \mathcal{H}$.

It turns out that in an RKHS, the reproducing kernel K is always symmetric and positive semi-definite, with the consequence that for any given set of points $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ from $\mathcal{X}$, the corresponding matrices $\boldsymbol{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$ are symmetric and positive semi-definite. One of the earliest uses of reproducing kernel Hilbert spaces in statistical function estimation can be traced back to the seminal work by Craven and Wahba (1979). For our purposes though, the most important result is the so-called representer theorem of Kimeldorf and Wahba (1971) of which a simplified version is given below, and for which a complete proof can be found in Wahba (1990).

Theorem 1.1. *Let $\mathcal{H} \subseteq C(\mathcal{X})$ be a reproducing kernel Hilbert space with reproducing K. Then, for every function $f \in \mathcal{H}$, problem (1.1) has a unique solution f_λ of the form*

$$f_\lambda = \sum_{j=1}^{n} \mathrm{w}_j K(\cdot, \mathbf{x}_j) \tag{1.2}$$

where the vector $\mathbf{w} = (\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_n)^\top$ can be found as a solution to

$$\arg \min_{\mathbf{w} \in \mathbb{R}^n} \left\{ \|\mathbf{y} - \boldsymbol{K}\mathbf{w}\|_2^2 + \lambda \mathbf{w}^\top \boldsymbol{K} \mathbf{w} \right\} \tag{1.3}$$

with the matrix $\boldsymbol{K}$ given by $\boldsymbol{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$.

The importance of the above theorem lies in the fact that not only does it provide an objective functional that leads to a unique estimator of f^*, but it also gives a way to actually compute that solution. Now, the representation depicted in (1.2) makes no specific assumption about the form of the kernel K. For the purposes of this paper, we shall consider *translation invariant radial* kernels, meaning that we will use kernels $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that there exists an even function $\phi : \mathbb{R}^+ \to \mathbb{R}$ such that

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\|\mathbf{x}_i - \mathbf{x}_j\|)$$

for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^p$. As a result, we will be approximating the unknown true function f^* using functions f_λ such that for a given $\mathbf{x} \in \mathcal{X}$,

$$f_\lambda(\mathbf{x}) = \sum_{j=1}^{n} \mathrm{w}_j \phi(\|\mathbf{x} - \mathbf{x}_j\|). \tag{1.4}$$

The real values $\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_n$ are known as the *weights* and the vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ are the *centers*. Function approximators of the form (1.4) became known in the Neural Networks and Machine Learning community as Radial Basis Function Networks where their popularity skyrocketed from the late 1980s to the entirety of the 1990s and even up until the present day, thanks in part to the so-called universal approximation theorem of Powell stated below, but maybe most importantly to their natural ability to provide approximator of smooth functions in high dimensional spaces. It is important however to note that the formulation of Radial Basis Function Networks assumes the existence of a fixed number k of centers, where $k \ll n$ is substantially less than the number n of training samples. In other words, a typical RBF network will be of the form

$$f_\lambda(\mathbf{x}) = \sum_{j=1}^{k} \mathrm{w}_j^* \phi(\|\mathbf{x} - \mathbf{x}_j^*\|). \tag{1.5}$$

where the centers $\mathbf{x}_1^*, \cdots, \mathbf{x}_k^*$ are p-dimensional vectors to be learned from the training data along with their corresponding weights $\mathrm{w}_1^*, \cdots, \mathrm{w}_k^*$. As far as the basis functions are concerned, the most popular choices include the so called Gaussian radial basis function kernel with $\phi(\mathrm{u}) = \exp(-r\mathrm{u}^2)$ where r represents the bandwidth. Some generalizations of the Gaussian RBF kernel use a different bandwidth for each of the attributes of the input vector $\mathbf{x}$. Other somewhat popular choices of kernels include: the multiquadrics kernel with $\phi(\mathrm{u}) = \left(r\mathrm{u}^2 + 1\right)^{\frac{1}{2}}$, the inverse multiquadrics kernel with $\phi(\mathrm{u}) = \left(r\mathrm{u}^2 + 1\right)^{-\frac{1}{2}}$ and the thin plate spline kernel for which $\phi(\mathrm{u}) = \mathrm{u}^2 \log(\mathrm{u})$. The universal approximation theorem for RBF networks can be stated as follows:

Theorem 1.2. *Let $\Phi : \mathbb{R}^p \to \mathbb{R}$ be a continuous and integrable function such that*

$$\int_{\mathbb{R}^p} \Phi(\mathbf{x}) d\mathbf{x} \neq 0.$$

Let $\mathcal{H}_\Phi$ denote the family of RBFs generated by Φ, i.e.

$$\mathcal{H}_\Phi = \left\{ f : f = \sum_{i=1}^{k} \mathrm{w}_i^* \Phi\left(\cdot - \mathbf{x}_i^*\right),\ k \in \mathbb{N}, \mathbf{x}_i^* \in \mathbb{R}^p, \mathrm{w}_i \in \mathbb{R} \right\}$$

Then, for any continuous function f^*, $\exists \tilde{f}_k \in \mathcal{H}_\Phi$ such that $\forall \epsilon > 0$,

$$\|\tilde{f}_k - f^*\| < \epsilon.$$

In other words, for any continuous function f^*, there exist (a) a kernel K along with the corresponding Φ, (b) an optimum number of basis functions k, (c) a set of weights $\{\mathrm{w}_i^*\}_{i=1}^k$ and (d) a set of centers $\{\mathbf{x}_i^*\}_{i=1}^k$ such that the corresponding function $\tilde{f}_k \in \mathcal{H}_\Phi$ approximates f^* to any desired precision ϵ.

Theorems (1.1) and (1.2) provide all the needed theoretical justification for using the large class of radial basis function approximators in regression. From a statistical model selection perspective, one could think of equation (1.4) as the deterministic portion of the full model while equation (1.5) would represent the deterministic portion of the optimal model once the most relevant atoms (bases) have been selected. It turns out, thanks to theorem (1.1), that the weights $\{\mathrm{w}_i\}_{i=1}^n$ of the full model can be estimated quite readily using traditional statistical estimation tools, once the suitable basis function is chosen. Indeed, equation (1.3) of theorem (1.1) clearly defines the regularized optimization to be solved in order to find the w_i's of the full model. For a suitable basis set however, it turns out to be more important to select the most relevant bases, just like one would want to select the most relevant variables in multiple linear regression. In other words, while a full solution is important, a much more parsimonious solution is more desirable for a variety of practical and theoretical reasons that will not be detailed in the present paper. It suffices to note that many researchers have proposed a variety of techniques for finding a sparse representation to equation (1.4), essentially constructing equation (1.5). From a modeling standpoint, it is important to note that theorem (1.1) is a simplified version of a much general theorem that allows the expansion to have a y-intercept captured by a different basis. However, our simplified version does not lose generality, since we assume throughout this paper that the data are standardized.

From a Bayesian perspective, equation (1.3) can be interpreted as the formulation a Maximum A Posteriori estimation of the vector $\mathbf{w}$, with a Gaussian likelihood given by $\mathbf{y} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{K}\mathbf{w}, \sigma^2 \mathbf{I}_n)$ and

a Gaussian prior given by $\mathbf{w} \sim \boldsymbol{\mathcal{N}}(0, (2\lambda \boldsymbol{K})^{-1})$. This prior corresponds to the so-called Silverman's g-prior for $\mathbf{w}$. It goes without saying that this formulation presupposes that the matrix $\boldsymbol{K}$ is invertible, which unfortunately in most situations turns out not to be the case, because of poor matrix conditioning, due either to the kernel, or to aspects of the data. At the very least however, the formulation provides an interesting starting point for considering a Bayesian perspective to the derivation of a sparse representation of equation (1.4). First of all, it is worth noting that the prior does not have to be Gaussian, and even when it turns out to be, the corresponding variance-covariance matrix does not have to be data-dependent as in the Silverman's g-prior case above. Given the Gaussian likelihood provided by $\mathbf{y} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{K}\mathbf{w}, \sigma^2 \mathbf{I}_n)$, we seek to specify a prior over $\mathbf{w}$ such that the Bayesian estimator $\tilde{\mathbf{w}}$ of $\mathbf{w}$ is k-sparse, i.e. has only k nonzero entries. This problem of sparse Bayesian learning in the context of radial basis function networks has been scrutinized by several authors: Tipping (2001) has introduced and developed the Relevance Vector Machine (RVM). The gist of the RVM approach lies in specifying an independent Gaussian prior for each w_i, each with a different precision hyperparameter λ_i. It then turns out that the use of a suitably specified Gamma distribution for each λ_i leads to marginal prior for the vector $\mathbf{w}$ that exhibits sparsity inducing contours. Practically, the computed RVM solution contains may λ_i that so large as to be considered infinite in magnitude, leading to the corresponding w_i being deemed to be essentially 0. RVM estimates are essentially obtained via Maximum A Posteriori. Perhaps it is worth pointing out that RVM is really a Bayesian generalized ridge regression with a suitable choice of hyperprior for the rigde parameters. Fokoué (2008) adopts a fully Bayesian approach to the same problem, essentially combining insights from Tipping (2001) and Barbieri and Berger (2004) to derive a model space search strategy for selecting the k most relevant basis elements from the original set of n provided by the full model. It turns out however that the raw output of the search procedure has potential of yielding unstable - in a sense that will be clarified later - estimates of the quantities of interest. In this paper, we propose a post-processing of the sample path that helps derive stable estimates. To help illustrate the point, let's assume that the underlying true function is

$$f^*(\mathbf{x}) = \mathrm{sinc}(\mathbf{x}) = \frac{\sin(\mathbf{x})}{\mathbf{x}}, \qquad \mathbf{x} \in [-10, 10].$$

Let's now generate data points $(\mathbf{x}_i, Y_i)$ with $Y_i = f^*(\mathbf{x}_i) + \epsilon_i$, and $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathbf{N}(0, 0.2^2)$. Finally, let's assume a radial basis function representation with the underlying kernel given by

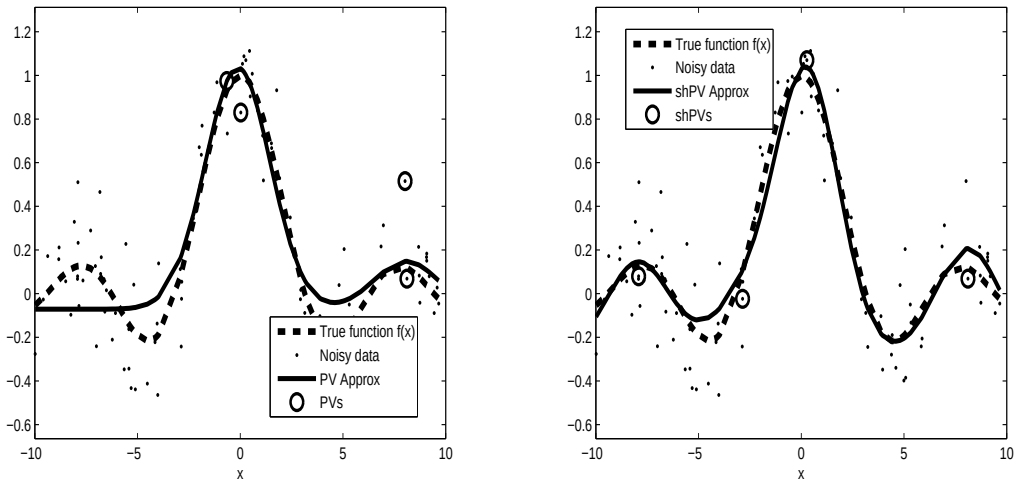
$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2r^2}\right).$$

The estimate of the length scale (bandwidth) r is found to be 2.2 for this dataset. The fit in Fig 1(b) was obtained after using mixture modelling to postprocess the sample path of the model search procedure. The fit is clearly better than the one in Fig 1(a) which was obtained by simply picking the k vectors with the largest prevalence. While Fokoué (2008) considered a variety of scenarios ranging from orthogonal polynomial regression to traditional multiple linear regression, this paper focuses solely on radial basis function regression. Following a notation introduced by Barbieri and Berger (2004) and used in Fokoué (2008), we start by defining the *full* model

$$\mathbf{y} = \boldsymbol{K}\mathbf{w} + \boldsymbol{\epsilon}, \tag{1.6}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^\top$ and $\boldsymbol{\epsilon} \overset{iid}{\sim} \boldsymbol{\mathcal{N}}(0, \sigma^2 \mathbf{I}_n)$. Both Barbieri and Berger (2004) and Fokoué (2008) considered selecting the *most predictively optimal* from among submodels of the form

$$M_{\mathbf{v}}: \quad \mathbf{y} = \boldsymbol{K}_{\mathbf{v}}\mathbf{w}_{\mathbf{v}} + \boldsymbol{\epsilon} \tag{1.7}$$



(a) Fit from raw sample path (b) Fit through mixture modelling

FIG 1. *Results of function approximation: (left) fit obtained using the raw sample path; (right) Fit obtained after post processing the sample path using mixture modeling*

where $\mathbf{v} = (v_1, \cdots, v_n)$ is the model index, defined coordinate-wise as follows:

$$v_i = \begin{cases} 1 & \textit{if the } i^{th} \textit{ column of } \boldsymbol{K} \textit{ is used by model } M_{\mathbf{v}} \\ 0 & \textit{otherwise} \end{cases} \tag{1.8}$$

In equation (1.7) above, $\boldsymbol{K}_{\mathbf{v}} \in I\!R^{n \times m}$ contains the columns of $\boldsymbol{K}$ corresponding to the nonzero coordinates of $\mathbf{v}$, and $\mathbf{w}_{\mathbf{v}} \in I\!R^m$ is the corresponding vector of regression coefficients. Here, $m = |\mathbf{v}|$ is simply the model size corresponding to the number of nonzero regression coefficients. Clearly, the model selection problem at hand is an old one. Both the statistics and machine learning literatures are rich with a wide variety of model selection techniques that have been invented and applied from both the frequentist and Bayesian perspectives. A distinct feature of both Barbieri and Berger (2004) and Fokoué (2008) that is worth emphasizing is that unlike the majority of authors before them, both these papers seek to select models for optimal prediction rather than model identification (explanation). Fokoué (2008) goes a little further by proposing a more flexible counterpart to the Median Probability Model (MPM) of Barbieri and Berger (2004), making it possible to avoid cases of non existing solutions (albeit at the cost of somewhat suboptimality) on the one hand and to handle cases of non full rank data matrix on the other hand. The reader is referred to the two papers for more detailed accounts of the techniques proposed therein. Despite providing a nice extension to its predecessor, Fokoué (2008) handling of non full rank data matrix cases produces model search results that tend yield rather unstable basis function selections. The idea proposed in this paper helps circumvent that difficulty. The rest of this paper is organized as follows: Section 2 provides a quick overview of the model selection technique used. Section 3 presents motivating examples and highlights the need for a better and more stable post processing of the sample path obtained from the model space search. Section 4 presents a simple mixture modeling approach to stabilizing the basis function selection, along with illustrations of

how the problem encountered earlier is solved using the proposed technique. Section 5 provides some discussion and conclusion.

2. Aspects of model space search

Our model selection approach is centered around the model index $\mathbf{v}$, since the entries of $\mathbf{v}$ indicate whether or not a particular atom is included in the model under consideration. From Bayesian perspective, this translates into the need to specify a prior on $\mathbf{v}$ and then deriving the corresponding posterior. Specifically, our approach is based on the overall posterior probability that an atom is included in a model out of $2^n - 1$ possible models.

Definition 1. *The posterior inclusion probability for atom i is*

$$p_i \equiv \sum_{j=1}^{2^n-1} I(\mathbf{v}_{ij} = 1) p(\mathbf{v}_j|\mathbf{y}) \tag{2.1}$$

p_i as defined above is nothing but $\Pr(v_i = 1|\mathbf{y})$, and represents the proportion of times atom i is chosen by one of the $2^n - 1$ models for predicting the response Y. It makes sense that the larger p_i is, the more important and therefore relevant atom i is. Obviously, no technique worth mentioning would dare to exhaustively search among $2^n - 1$ models, even for the smallest of samples. In other words, we do not explicitly calculate each p_i based on $2^n - 1$: as remarked by Barbieri and Berger (2004), it suffices to estimate the p_i from the sample path yielded by the model space search as we explain in details in the following section. First of all, we use a noninformative prior for each atom, namely

$$\Pr[\mathrm{v}_j = 1] = \frac{1}{2},\ j = 1, \cdots, n,$$

so that each model index arises with equal probability $p(\mathbf{v}) = \frac{1}{2^n}$, regardless of its size. As for the model size k, it is implicitly tied to the model index. The model search strategy that we use is based on the simulation of a continuous-time birth-and-death process. Specifically, we consider a set $\boldsymbol{\mathcal{A}}$ consisting of the indices of those atoms that make up the current model. We then allow new atoms to be added to $\boldsymbol{\mathcal{A}}$ or atoms to be removed from $\boldsymbol{\mathcal{A}}$ based on their contribution to the marginal likelihood. With $p(\mathbf{y}|\sigma^2, \lambda, \mathbf{v})$ representing the marginal likelihood associated with model $\mathbf{v}$, and $\mathbf{v}\backslash\{i\}$ representing the $(|\mathbf{v}| - 1)$-model without atom i, the death rate δ_i for atom i is given by

$$\delta_i \propto \frac{p(\mathbf{y}|\sigma^2, \lambda, \mathbf{v}\backslash\{i\})}{p(\mathbf{y}|\sigma^2, \lambda, \mathbf{v})}.$$

Once the δ_i's are computed for all the atoms, the overall death rate $\delta = \sum_{i=1}^{|\mathbf{v}|} \delta_i$ is computed to determine whether a birth or death needs to occur. We use an overall constant birth rate ν, so that a birth occurs according to a Bernoulli draw with parameter $\nu/(\nu + \delta)$. With

$$\eta \sim \mathsf{Ber}\left(\nu/(\nu + \delta)\right),$$

the next event is a birth if $\eta = 1$. Within the birth-and-death process, an event is either a birth or a death, and the time to the next event is assumed to be exponentially distributed with parameter $1/(\nu + \delta)$,

$$t \sim \mathsf{Exp}\left(1/(\nu + \delta)\right).$$

Starting at 0, one sweep of the birth-and-death process runs for a total T units of time with each increment drawn from an exponential. This use of the exponential distribution for the time between consecutive events dovetails with our use of a truncated Poisson distribution as our prior for the size of the model. In this paper, we consider the simplest of prior over the weights, namely the isotropic Gaussian prior $\mathbf{w} \sim \mathcal{N}(0, \lambda^{-1}\mathbf{I}_n)$. Although this prior does not inherently have the ability to yield sparse estimates of $\mathbf{w}$, the fact that we search a large space of models of varying size does not hinder our ability to select the atoms that yield optimal prediction. With this prior, the much needed marginal likelihood for a submodel $M_{\mathbf{v}}$ is Gaussian and given by,

$$p(\mathbf{y}|\sigma^2, \lambda, \mathbf{v}) = \frac{1}{\sqrt{(2\pi)^n \det(\sigma^2\mathbf{I}_n + \lambda \boldsymbol{K}_{\mathbf{v}}\boldsymbol{K}_{\mathbf{v}}^{\mathsf{T}})}} \exp\left[-\frac{1}{2}\mathbf{y}^{\mathsf{T}}(\sigma^2\mathbf{I}_n + \lambda \boldsymbol{K}_{\mathbf{v}}\boldsymbol{K}_{\mathbf{v}}^{\mathsf{T}})^{-1}\mathbf{y}\right]$$

Putting all the above together, a pseudo-code for our model space search is given below.

Algorithm: Model space search

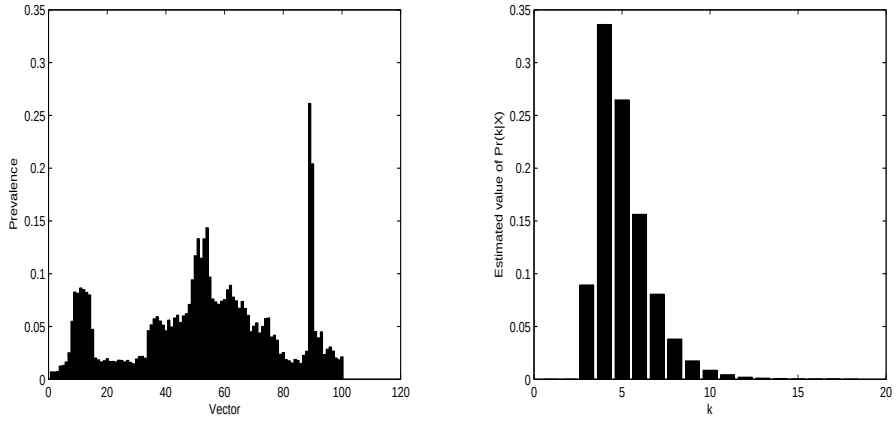
- Initialize $\mathcal{A}^{(0)}$ with $\lfloor n/2 \rfloor$ indices randomly drawn from $\{1, 2, \cdots, n\}$.
- Initialize posterior inclusion probabilities: $p_j = 0, \quad j = 1, \cdots, n$.
- Repeat
  - $r := r + 1$
  - Update the active set
    $\mathcal{A}^{(t)} :=$ birth-and-death$(\mathcal{A}^{(r-1)}, \lambda^{(r-1)}, (\sigma^2)^{(r-1)}, \mathbf{y})$
  - Update the posterior inclusion probabilities
    for $j := 1$ to n
      if $j \in \mathcal{A}$ then $p_j := p_j + 1$
    end
    $p^{(r)} := (p_1, p_2, \cdots, p_m)$
  - Get new model size
    $k^{(r)} := |\mathcal{A}^{(r)}| :=$ length$(\mathcal{A}^{(r)})$
  - Estimate current parameters
    $(\lambda^{(r)}, (\sigma^2)^{(r)}) :=$ Gibbs-sampling$(\lambda^{(r-1)}, (\sigma^2)^{(r-1)}, \mathcal{A}^{(r)}, \mathbf{y})$

Until $r = R$

More details about the birth-and-death process simulation can be found in Fokoué (2008).

3. Mixture Modeling of the Sample Path for Stability

Let's consider once again the motivating example of the *sinc* function presented earlier. If we use our model space search strategy for this example, the corresponding sample path allows us to obtain approximate distributions for both the model size k and the model index $\mathbf{v}$ as shown on Figure (2). Figure (2(b)) strongly suggests $k = 4$ or $k = 5$ as the most plausible candidates for optimal model size. We then consider picking the 4 or 5 atoms with the highest posterior inclusion probabilities. However, due to good mixing and neighborhood effect, the 4 atoms directly selected from Figure (2(a)) suffer from what we call the "redundancy of potential prevalent atoms". Indeed, two of the highest values of p_j correspond to two vectors that are virtually equal in magnitude. Hence the



(a) Distribution of prevalence (b) Distribution of model size

FIG 2. *Approximate distribution of atom prevalence and approximate distribution of model size*

lesser quality of fit seen on Fig 1(b) of our motivating example. This redundancy is indeed to be expected when the raw sample path is used, partly because of the fact that the representation of the function is in dual space with weights being direct functions of the observations which in turn can be very close to one another. When such situations arise, they cause k-variable models to select all or most of the k prevalent atoms in the same neighborhood. The problem is exacerbated when the number of model space searches is made very large as shown in Fig 2(a). One could argue that this situation is an artifact of the search technique, and that a solution consists in redesigning the search technique. While that may be a viable solution, we feel that for a well-mixing search algorithm like ours, it might be better to seek a solution that leaves the search algorithm intact, and instead post processes the output. That's exactly the contribution of this paper. Noticing that the plot of the distribution of the prevalence of atoms exhibits a multi-modal shape, it seemed natural to extract the modes of this distribution and use those modes as the *prevalent points*. The sample path of the model search procedure contains a sequence $\{p_i,\ i = 1, \cdots, p\}$ where p_i represents the number of times atom i was in the current model during the R searches of the model space. Fig 2(a) is a plot of the histogram of such a sample path. Clearly, its shape suggests the plausibility of mixture modeling as a way to extract the modes. Let's consider $\{p_i,\ i = 1, \cdots, p\}$ where p_i represents the number of times atom i was in the current model during the R searches of the model space. The key idea consists in forming a new sample of size R from the sample path. One could think of it as drawing a stratified sample of sorts, with each $\mathbf{x}_i$ being replicated p_i times to reflect its prevalence. It actually turns out that one could form the new sample $\boldsymbol{S}$ during the model space search by updating it after each sweep of the simulation of the birth-and-death process. Basically, having initialized $\boldsymbol{S}$ as an empty set, augment it after each sweep of the process using

$$\boldsymbol{S} := \boldsymbol{S} \cup \{\mathbf{x}_j\},$$

for all the atoms currently in the active set, i.e., for all j such that $j \in \boldsymbol{\mathcal{A}}$. Alternatively, one could post-process the sample path as follows:

Algorithm: Postprocessing of the sample path

Initialize sample $\boldsymbol{S}$ as empty
for $i = 1$ to n
$\boldsymbol{S} := \boldsymbol{S} \cup \{p_i \text{ copies of } \mathbf{x}_i\}$.
end

One can think of this operation as reconstructing the original sample from a stem and leaf plot. We now treat $\boldsymbol{S}$ as a random sample from the distribution of a random variable U. Using the fact that the overall search scheme also produces the most probable number of prevalent atoms k^*, and considering the fact the above approximate distribution of the indices exhibits strong multimodality, it is reasonable to use a mixture of k^* Gaussians to model the distribution of U, namely

$$\boldsymbol{p}(u) = \sum_{\ell=1}^{k^*} \pi_\ell \boldsymbol{p}_\ell(u; \mu_\ell, \sigma_\ell^2)$$

An implementation of the standard EM algorithm for mixtures can then be used to find estimates of the centers $\hat{\mu}_1, \cdots, \hat{\mu}_{k^*}$. Now, for $j = 1, \cdots, k^*$, the prevalent vectors $\{\mathbf{x}_1^*, \cdots, \mathbf{x}_{k^*}^*\}$ are given by

$$\mathbf{x}_j^* = \arg \min_{\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}} \|\mathbf{x}_i - \hat{\mu}_j\|^2,$$

from which the corresponding fit is given by

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{k*} \hat{\mathrm{w}}_j^* K(\mathbf{x}, \mathbf{x}_j^*)$$

3.1. Example of fit using mixture modelling

Let's consider once again the underlying true function

$$f(\mathbf{x}) = \mathrm{sinc}(\mathbf{x}) = \frac{\sin(\mathbf{x})}{\mathbf{x}},$$

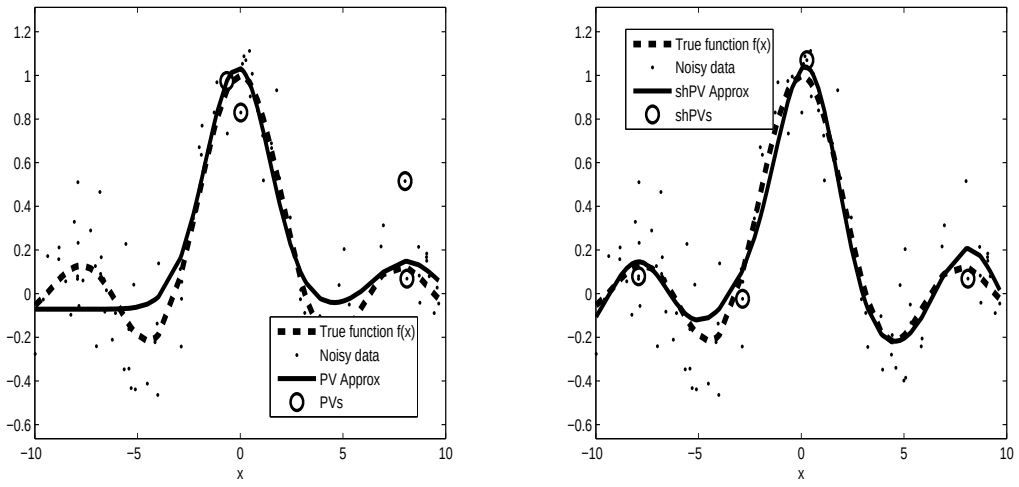
with data points $(\mathbf{x}_i, Y_i)$ assumed to arise from the representation

$$Y_i = \sum_{j=1}^{n} \mathrm{w}_j^* B_j(\mathbf{x}_i) + \epsilon_i,$$

with $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathbf{N}(0, 0.2^2)$ and basis function given by

$$B_j(\mathbf{x}_i) = K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2r^2}\right).$$

The estimate of the length scale (bandwidth) r is found to be 2.2 for this dataset. The fit in Fig 3(b) was obtained after using mixture modelling to postprocess the sample path of the model search procedure. The fit is clearly better than the one in Fig 3(a) which was obtained by simply picking the k vectors with the largest prevalence.



(a) Fit from raw sample path (b) Fit through mixture modelling

FIG 3. *Results of function approximation: (left) fit obtained using the raw sample path as is (right) Fit obtained after post processing of the sample path using mixture modeling*

4. Conclusion and discussion

We have proposed a straightforward and effective solution to an important aspect of model space search in the context fully Bayesian treatment of radial basis function regression. In our future work, we intend to enrich our search algorithm in such a way to sharpen the sample path to clearly isolate the modes even in the presence of the desirable good mixing. We intend to run our proposed algorithm on a variety of high dimensional simulated and real life data. We also plan on exploring the effect of a different prior distribution on the weights of the radial basis function expansion. Of particular interest will be the use of Silverman's g-prior along the lines of Zhang et al. (2008) who theoretically proved the consistency of the posterior when Bayesian model selection is tackled using Silverman's g-prior. In the same spirit, we will consider exploring the effect of Zellner's g-prior, following the recent work by Liang et al. (2008).

References

Barbieri, M. and J. O. Berger (2004). Optimal predictive model selection. *Ann. Statist. 32*, 870–897.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of GCV. *Numer. Math 31*, 377–403.

Fokoué, E. (2008). Estimation of atom prevalence for optimal prediction. *Contemporary Mathematics 443*, 103–129.

Kimeldorf, G. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic. 33*, 82–85.

Liang, F., R. Paulo, G. Molina, M. Clyde, and J. O. Berger (2008). Mixtures of g-priors for Bayesian variable selection. *J. Amer. Statist. Assoc. 103*, 410–423.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *J. Mach. Learning Res. 1*, 211–244.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59. Philadelphia: SIAM CBMS-NSF Regional Conference Series.

Zhang, Z., M. I. Jordan, and D. Yeung (2008). Posterior consistency of the silverman g-prior in bayesian model choice. Technical report, University of California, Berkeley, California, USA, Department of Electrical Engineering and Computer Science.