Bayesian Additive Regression Kernels

Zhi Ouyang  
Google, Inc.  
Mountain View, CA 94043, USA  

Merlise A. Clyde  
clyde@stat.duke.edu  
Robert L. Wolpert  
wolpert@stat.duke.edu  

Department of Statistical Science  
Duke University  
Durham, NC 22208-0251, USA  

Editor: tbd

Abstract

We propose a general Bayesian “sum of kernels” model, named Bayesian Additive Regression Kernels (BARK), for regression and classification problems. The unknown mean function is represented as a weighted sum of kernel functions, constructed using a prior built from $\alpha$-Stable Lévy random fields. This leads to a specification of a joint prior distribution for the number of kernels, kernel regression coefficients and kernel location and scale parameters. We show that the $\alpha$-Stable prior on the kernel regression coefficients may be approximated by $t_\alpha$ distributions, leading to efficient inference via a scale mixtures of normal representation. With a heavy tail prior distribution on the kernel regression coefficients and a finite support on the kernel location parameters, BARK achieves sparse representations. The scale parameters in the kernel functions allow feature selection of predictors and capture non-linear interactions of the variables. A reversible-jump Markov chain Monte Carlo algorithm is developed to make posterior inference on the unknown mean function. We show how the Gaussian model may be extended to binary regression/classification by the introduction of latent variables. We compare the method to support vector machines and BART (Bayesian Additive Regression Trees) on simulated and real data sets.

Keywords: Supervised learning, Feature Selection, Lévy random field

1. Introduction

In supervised learning we are given a sample of size $n$ of observed input vectors $\{x_i\}_{i=1}^n$, $x_i \in \mathcal{X}$ and responses $\{y_i\}_{i=1}^n$ with dual goals of discovering relationships between the input and the response variables from a training data set, and then making accurate predictions of $Y$ for new values of $x$ in a held-out sample. Often the response variable $Y = (y_1, \ldots, y_n)^T$ is one dimensional, either continuous (as in regression) or discrete (as in classification), while the covariates are multidimensional, say $x_i \in \mathbb{R}^p$, where $p$ may in fact be much greater than $n$.

For continuous responses, nonparametric regression models provide a flexible and adaptive approach to model the relationship between the response $Y_i$ and covariates $x_i$. A popular candidate for the mean function $E[Y \mid x] = f(x)$ is constructed by the sum of
generating functions,

\[ f(x) = \sum_j g(x, \theta_j)\beta_j, \]  

where \( \theta_j \) is the parameter vector of the \( j \)th generating function and \( \beta_j \) is a weight. For example, neural networks (Cheng and Titterington, 1994) often use logistic functions for \( g \). Bayesian Additive Regression Trees (BART) (Chipman et al., 2007) uses tree models as the generating functions. Alternatively, kernel methods have been extensively studied in the machine learning and statistics literature (Hofmann et al., 2008; Pillai et al., 2007) for classification and prediction, where the kernel function \( k \) is a positive definite function mapping \( \mathcal{X} \times \mathcal{X} \) to \( \mathbb{R} \), satisfying \( k(x, x') = \langle \Phi(x), \Phi(x') \rangle \) for all \( x, x' \in \mathcal{X} \). The function \( \Phi \) maps the input vectors \( x \) into the “feature” space, permitting dimension reduction of the original input vectors \( x \). Neural networks, “sum-of-trees” and “sum-of-kernels” models explore the additive effects through the linear combination of the different generating functions, and explore the interactive and nonlinear effects through the individual non-linear generating functions.

For kernel models, such as the well-known Support Vector Machine (SVM) (Boser et al. 1992 Cristianini and Shawe-Taylor 2000, Chap. 6), the standard approach is find a function \( f \in \mathcal{H}_k \) in a reproducing kernel Hilbert space (RKHS) that is the solution to minimizing a penalized loss function of the form

\[ L(f; Y) + \lambda \| f \|_{\mathcal{H}_k}^2 \]  

where the first term captures the closeness of the function to the data (often induced by a log-likelihood from the sampling distribution of the data), while the second term is a smoothness penalty induced by the norm of the function in the RKHS; \( \lambda \) is a penalty that balances goodness of fit with model complexity. The representer theorem (Kimeldorf and Wahba, 1971) leads to a finite dimensional solution which uses \( n \) kernels

\[ \hat{f}(x) = \sum_{j=1}^{n} k(x, x_j)\hat{\beta}_j. \]

supported at observed input vectors. While the solution involves \( n \) kernels, the penalized loss function permits over-fitting. Using properties of the RKHS, the finite dimensional penalized loss function maybe expressed as

\[ L(f; Y) + \lambda \beta^T K \beta \]  

where \( \beta = (\beta_1, \ldots, \beta_n)^T \) is the vector of regression coefficients and \( K = [k(x_i, x_j)] \) is the \( n \times n \) kernel matrix. From a Bayesian perspective, the penalty may be viewed as being induced by a Gaussian prior distribution, where \( \beta \sim N(0, \lambda^{-1} K^{-1}) \). Relevance Vector Machines (RVM) (Tipping, 2001) also uses \( n \) kernels centered at training samples but seek to find a sparser solution than SVM by employing an alternative penalty. In the regression case, RVM assumes a Gaussian additive error,

\[ Y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \]
and maximizes a type-II likelihood under following prior distributions,

\[ \beta_j \overset{\text{iid}}{\sim} \text{No}(0, \varphi_j^{-1}), \quad \varphi_j \overset{\text{iid}}{\sim} \text{Ga}(a, b), \quad 1/\sigma^2 \sim \text{Ga}(c, d). \quad (5) \]

under the limiting case where \( a = b = c = d = 0 \).

Both SVM and RVM provide accurate predictions/classifications, but only deliver a point estimator. A fully Bayesian approach has the advantage of making probabilistic statements for predictions and permits inference about associated kernel parameters for feature selection. While the hierarchical framework in RVM seems to suggest that a fully Bayesian solution is possible, the recommended hyperparameters \( a = b = c = d = 0 \) lead to an improper posterior distribution for \( f \). While Chakraborty et al. (2012) developed fully Bayesian versions of SVM and RVM using proper prior distributions, their solution (as well as Tipping) is applied to the finite representation given by the solution from the representer theorem, rather than placing prior distributions directly on functions in some function space. By considering prior distributions directly on the function space, we may relax the limitations of the RKHS approach.

In this paper, we detail another fully Bayesian framework for supervised learning with mean function Equation (1), a.k.a. Bayesian Additive Regression Kernels (BARK). Instead of using a fixed number of kernel functions, as in SVM, RVM and the Bayesian counterpart of these models, we allow the number of kernel functions to be random. Conditional on the number of kernels \( J \), adopt similar prior distributions for the regression coefficients as in RVM Equation (5). When \( b \) goes to zero, the posterior distribution for \( \beta \) is improper if \( J \) is fixed. One way to overcome this impropriety problem is to specify the prior distributions for \( \beta \) and \( J \) jointly. A small \( b \) yields a large \( \varphi \) and a small \( \beta \). Since the mean function \( f(x) \) is constructed by the sum of these kernels, if each of the regression coefficients is small, it needs a large amount of small kernels to re-build the mean function on the same scale. Allowing the number of kernels goes to infinity while the regression coefficients shrinks to zero defines the prior distributions consistently. [wow.] In the limit, the mean function can be viewed as the sum of infinitely many tiny kernel functions. This prior distribution becomes an infinite divisible random field with independent increments, or Lévy random field in the limit. Lévy random field has already been used in kernel regression problems, such as Clyde et al. (2005, 2006); Clyde and Wolpert (2007); Woodard et al. (2010); House et al. (2011); Wolpert et al. (2011), and we shall extend this approach to supervised learning problems in this paper.

We introduce a generalization of independent Cauchy prior distributions for nonparametric regressions, which is called symmetric \( \alpha \)-Stable /Lévy random field. It induces a heavy polynomial tail for the prior distribution on regression coefficient, which favors a sparse representation in the model. In practice, we need to approximate the random measure due to computing limitations, but the theory guarantees consistency when the approximation approaches the true measure, hence we have a valid full Bayesian specification in the limit. We extend this approach to the classification problems, which is the first time to apply the Lévy random field theory in this scenario.

Most kernel regression models only focus on the learning of kernel location parameters, but not the kernel scale parameters. For example, the original SVM, the hierarchical Bayes SVM and RVM in (Chakraborty et al., 2012) uses kernels with a single scale parameter, such as the Gaussian kernel whose precision matrix is a scale multiplied by the identity
matrix. These kernels assume homogeneity across all covariates, which is usually not true in modern applied problems, particularly when the number of covariates $p$ is large. We use Gaussian kernels with diagonal precision matrix, which assigns a scale parameter for each covariate. This facilitates a variety of structures that can be used in feature selection. Similar to the approach described in George and McCulloch (1997), we incorporate the hierarchical mixture prior distribution of a point mass at zero and a continuous distribution for the kernel scale parameters to enable selection process.

In the next section, we present the details of BARK using symmetric $\alpha$-Stable Lévy random field as the prior distribution. We describe the prior distributions on the kernel location parameters that induces sparse representations. We detail four different settings for the kernel scale parameters so that feature selection can be achieved under different scenarios. We explain how to elicit the hyperparameters, and how marginalization can be used to make the Markov chain mix faster. The framework is then extended straightforwardly to the classification problems in Section 3. We demonstrate the approach through several simulated and real data sets in Section 4, and concludes in Section 5.

2. Bayesian Additive Regression Kernels

The BARK model for Gaussian responses $\{y\}_{i=1}^n$ and covariate vectors $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$ takes the form:

$$y_i = f(x_i) + \epsilon_i$$

for independent additive Gaussian errors $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \phi^{-1})$ with precision $\phi > 0$, and mean function expressed (under regularity conditions described below) as a weighted sum of kernels

$$f(x) = \sum_j \beta_j k(x, \chi_j)$$

with finitely- or infinitely-many random coefficients $\{\beta_j\} \subset \mathbb{R}$, random variables $\{\chi_j\}$ taking values in some complete separable metric space $\mathbb{X}$, and kernel $K : \mathbb{R}^p \times \mathbb{X} \rightarrow \mathbb{R}$. The mean function can also be represented as a stochastic integral

$$f(x) = \int_{\mathbb{X}} k(x, \chi) \mathcal{L}(d\chi),$$

with respect to the random signed Borel measure $\mathcal{L}(d\chi) = \sum_j \beta_j \delta_{\chi_j}(d\chi)$ which puts mass $\beta_j \in \mathbb{R}$ at location $\chi_j \in \mathbb{X}$. When $\{\beta_j, \chi_j\}$ are iid and their number $J$ is either infinite or Poisson-distributed, these are special cases of the LARK models of (Wolpert et al., 2011), where $\mathcal{L}(d\chi)$ is a random signed measure with a Lévy distribution, $\mathcal{L} \sim \mathcal{L}_\nu(\nu)$. The control measure $\nu$ determines the distributions of the $\{(\beta_j, \chi_j)\}_{0 \leq j < J}$ and their number $J \leq \infty$. A defining characteristic of $\mathcal{L} \sim \mathcal{L}_\nu(\nu)$ is its assignment of independent infinitely-divisible distributed random variables $\mathcal{L}(A_j)$ to disjoint Borel sets $A_j \subset \mathbb{X}$. In this paper we will focus on a particular example of infinitely-divisible distribution, the Symmetric $\alpha$-Stable (SoS). In Section 2.2 we show how the Cauchy distribution, the special case $\alpha = 1$ of the SoS family, offers significant computational advantages.
2.1 Symmetric $\alpha$-Stable Lévy Random Fields

The SoS random measure $\mathcal{L}(d\chi)$ has power-law control measure

$$\nu(d\beta, d\chi) = \gamma c_\alpha |\beta|^{-1-\alpha} d\beta \pi_\chi(d\chi)$$

for some intensity $\gamma > 0$, stable index $\alpha \in (0, 2)$, constant

$$c_\alpha = \frac{\alpha}{\pi} \Gamma(\alpha) \sin \left( \frac{\pi \alpha}{2} \right),$$

and Borel probability measure $\pi_\chi(d\chi)$ on $\mathbb{X}$. The random measure $\mathcal{L}(d\chi)$ assigns independent $\alpha$-Stable random variables $\mathcal{L}(A_j) \overset{\text{iid}}{\sim} \text{St}(\alpha, 0, \gamma \pi_\chi(A_j), 0)$ to disjoint Borel sets $A_j \subset \mathbb{X}$ (see Samorodnitsky and Taqqu 1994, p. 118; we use parametrization $(M)$ of Zolotarev 1986, p. 11).

Wolpert et al. (2011) show how to use a Poisson representation of Lévy random measures to support posterior Bayesian inference. Since the SoS control measure is infinite, the sum in Equation (7) will have infinitely-many terms. Only for $\alpha \in (0, 1)$ is the sum absolutely convergent; for $\alpha \geq 1$ (including the Cauchy case we consider further below) “compensation” is required and convergence is more delicate (Sato, 1999, Ch. 3). Fortunately in the symmetric case we consider here, where the measure $\nu(d\beta d\chi)$ is even in $\pm \beta$, the additional terms that compensation entails may be omitted (see Wolpert et al., 2011, p. 38 for details).

One approach to approximating the infinite sum in Equation (7) with a finite one is simply to truncate at some fixed threshold $\epsilon > 0$, including only those terms for which $|\beta_j| \leq \epsilon$. This leads to a similar sum with a Poisson-distributed number $J_\epsilon < \infty$ of terms. Wolpert et al. (2011) verify convergence (pointwise and in several function spaces) as $\epsilon \to 0$ and offer bounds on the error introduced by truncation. The conditional (given $J_\epsilon$) distribution of the coefficients $\beta_j$ are two-sided Pareto

$$\{\beta_j\}_{0 \leq j < J_\epsilon} \overset{\text{iid}}{\sim} \frac{\epsilon^\alpha}{2\alpha} |\beta|^{-1-\alpha} 1_{\{\beta > \epsilon\}} d\beta,$$

independent of $\{\chi_j\}_{0 \leq j < J_\epsilon} \overset{\text{iid}}{\sim} \pi(d\chi)$. Although truncation enables Bayesian inference for SoS random fields, the hard cut-off in truncating $\beta$ leads to unsatisfactory mixing of the Markov chain in practice. As an alternative, we here approximate the SoS random field by a Lévy field with Lévy measure:

$$\nu_\epsilon(d\beta, d\chi) = \frac{\gamma c_\alpha}{(\beta^2 + \alpha \epsilon^2)^{(\alpha+1)/2}} d\beta \pi_\chi(d\chi),$$

with finite mass

$$\nu_\epsilon^+(\alpha, \gamma) = \frac{\gamma \alpha^{1-\alpha/2}}{2^{1-\alpha} \epsilon^{\alpha}} \frac{\Gamma(\alpha/2)}{\Gamma(1-\alpha/2)}$$

or, for the Cauchy case, $\nu_\epsilon^+(1, \gamma) = \gamma/\epsilon$. This continuous approximation induces a Poisson-distributed finite number $J_\epsilon \sim \text{Po}(\nu_\epsilon^+(\alpha, \gamma))$ of terms, now with scaled central Student $t$-distributed coefficients

$$\{\beta_j\}_{0 \leq j < J_\epsilon} \overset{\text{iid}}{\sim} t_\alpha(0, \epsilon)$$
with α degrees of freedom, with density function
\[
\frac{\Gamma \left( \frac{\alpha+1}{2} \right)}{\Gamma(\alpha/2)(\alpha \pi)^{1/2}} \left( 1 + \frac{\beta^2}{\alpha \epsilon^2} \right)^{-\frac{\alpha+1}{2}}
\]

As \( \epsilon \to 0 \), the random variable \( f_\epsilon(x) \) in Equation (8) constructed from the approximate Lévy random measure converges to that \( f(x) \) without approximation. We prove this with a coupling argument, constructing both approximations \( \mathcal{L}_\epsilon \) and the limiting random measure \( \mathcal{L} \) on the same probability space:

**Theorem 1** Let \( \nu(d\beta, d\chi, du) = \gamma c_\alpha |\beta|^{-1-\alpha} d\beta \pi_\chi(d\chi) du \) be a Lévy measure on \( \mathbb{R} \times \chi \times (0, 1) \), where \( \gamma > 0 \), \( c_\alpha = (\alpha/\pi) \Gamma(\alpha) \sin(\alpha \pi/2) \) and \( \pi_\chi(d\chi) \) is a probability measure on \( \chi \). It induces a Lévy random field \( \mathcal{L} \) that maps a function \( g \in L_2(\chi, \pi_\chi(d\chi)) \) to the random variable

\[
\mathcal{L}[g] = \int_{\mathbb{R} \times \chi \times (0, 1)} (\beta - \sin \beta) g(\chi) \mathcal{N}(d\beta, d\chi, du) + \int_{\mathbb{R} \times \chi \times (0, 1)} \sin \beta g(\chi) \tilde{\mathcal{N}}(d\beta, d\chi, du)
\]

(12)

where

\[
\mathcal{N} \sim \text{Po}(\nu), \quad \tilde{\mathcal{N}}(d\beta, d\chi, du) = \mathcal{N}(d\beta, d\chi, du) - \nu(d\beta, d\chi, du).
\]

Then \( L[g] \sim \text{St}(\alpha, 0, \gamma^*, 0) \) with \( \gamma^* = \gamma \int_\chi |g(\chi)|^\alpha \pi_\chi(d\chi). \)

For any \( \epsilon > 0 \), construct the approximate Lévy random field \( \mathcal{L}_\epsilon \) that maps any function \( g \in L_2(\chi, \pi_\chi(d\chi)) \) to the random variable

\[
\mathcal{L}_\epsilon[g] = \int_{\mathbb{R} \times \chi \times (0, 1)} \beta g(\chi) 1_{\{ u < (1 + \alpha \epsilon^2 \beta^2 - \alpha)/2 \}} \mathcal{N}(d\beta, d\chi, du).
\]

(13)

Then \( \| \mathcal{L}_\epsilon[g] - \mathcal{L}[g] \|_2 \to 0 \) as \( \epsilon \to 0 \), for any \( g \in L_2(\chi, \pi_\chi(d\chi)) \).

The proof is given in Appendix A, and the squared discrepancy of the continuous approximation is finite:

\[
\mathbb{E} \left[ \| \mathcal{L}[g] - \mathcal{L}_\epsilon[g] \|_2^2 \right] \leq \|g\|^2 \frac{2\gamma}{\pi} \Gamma(\alpha + 1) \sin \left( \frac{\alpha \pi}{2} \right) \left( \frac{(1 + \alpha)|\alpha|/2}{2} + \frac{\alpha(2-\alpha)/2}{2 - \alpha} \right) \epsilon^{2-\alpha}.
\]

(14)

In particular, for the Cauchy the squared discrepancy can be calculated exactly when \( \alpha = 1 \):

\[
\mathbb{E} \left[ \| \mathcal{L}[g] - \mathcal{L}_\epsilon[g] \|_2^2 \right] = \int_{\mathbb{R} \times \chi} g(\chi)^2 \frac{2\gamma}{\pi} \left( 1 - \frac{1}{1 + (\epsilon/\beta)^2} \right) d\beta d\chi = \gamma \epsilon \|g\|^2_2.
\]

(15)

This offers guidance on the choice of parameters \( \gamma \) and \( \epsilon \) (see Section 2.4).
2.2 Sparse Representation

In this section, we shall detail the remaining prior distributions for BARK Equation (7) that obtains a sparse representation while selecting the features from the original covariate space.

Denote by $X$ the support set for kernel location parameter $\chi$. One possible decision is to set $X = \mathbb{R}^p$, and the kernel functions can be centered at any point in $\mathbb{R}^p$. This would lead to a flexible model, but the computation is demanding for large $p$ problems. On the other hand, we could continue the idea of SVM and RVM, whose kernels sit on observed data points, i.e., $X = \{x_1, x_2, \ldots, x_n\}$. This reduced the space $\mathbb{R}^p$ to $n$ discrete points.

Let $\pi(\chi) \sim \text{Un}(\{x_0, x_1, \ldots, x_n\})$. In order to incorporate the intercept term in the regression seamlessly into this representation, we add an imaginary point $x_0$ to $X$, such that $X = \{x_0, x_1, x_2, \ldots, x_n\}$, with $k(x, x_0) = 1$. It is natural to set the prior distribution for $\chi$ be uniformly over the set of possible kernel locations $\chi_j$ iid $\sim \text{Un}(\{x_0, x_1, \ldots, x_n\})$. (16)

This prior specification guarantees that there are at most $n + 1$ distinct kernels in the model Equation (7). When $\epsilon$ goes to zero, the prior distribution induces more and more kernels in the expression Equation (7), many of which will share the same location parameter. It is equivalent to the representation with unique kernels whose regression coefficients are the sum of all coefficient with the same kernel parameters. Suppose there are $n_i$ kernels centered at location $x_i$, and $J = \sum_{i=0}^n n_i$. The regression mean function can be rewritten as

$$f(x) = \sum_{i=0}^n \tilde{\beta}_i k(x, x_i), \quad \tilde{\beta}_i = \sum_{\{j \mid \chi_j = x_i\}} \beta_j.$$  

In particular, if $\alpha = 1$, not only the Cauchy process is infinite divisible, the approximated Cauchy prior distributions on the regression coefficients are also infinitely divisible. Therefore, the prior specification in Equations (11,16) becomes

$$J \sim \text{Po}(\nu^+_\epsilon(\alpha, \gamma, \epsilon))$$

$$n \mid J \sim \text{MN}(J, 1/(n + 1))$$

$$\{\tilde{\beta}_i\}_{i=0}^{n_i} \mid n \sim \text{Ca}(0, n_i\epsilon)$$  

(17)

where $n = (n_0, n_1, \ldots, n_n)$. As a result, kernels with the same location collapse, and this yield a sparse representation even when $\epsilon$ is small.

2.3 Feature Selection

In this paper, we focus on Gaussian kernels with diagonal covariance matrix, i.e.,

$$k(x, \chi) = \exp \left\{ - \sum_{l=1}^p \lambda_l (x_l - \chi_l)^2 \right\},$$  

(18)

where the scale parameters $\lambda_l$ $\forall l$ measure the contribution of the $l$th variable in the kernel function. We standardize each covariate before the analysis, i.e., variable $X_l$ has mean 0
and standard deviation 1 among the training samples. If $\lambda_l$ is zero, there is no contribution made by the $l$th variable through the kernel function; on the other hand, if $\lambda_l$ is large, the $l$th variable is important in the kernel regression.

We demonstrate four possible prior specifications for the scale parameters in the kernel function, BARK with equal weights, BARK with different weights, BARK with selection and equal weights, BARK with selection and different weights. The sum of the scale parameters in those four settings have the same prior distribution, which keeps the kernel function Equation (18) roughly in the same range since all variables are standardized. BARK with equal weights does not make any feature selection, BARK with different weights makes feature selection through soft shrinkage only, BARK with selection and equal weights makes feature selection through hard shrinkage only, while BARK with selection and different weights can make feature selection through both soft and hard shrinkage.

2.3.1 BARK with equal weights

The simplest kernel structure is to set the scale parameters $\lambda_l$s all equal. Suppose the prior distribution for the sum of all $\lambda_l$s is $\text{Ga}(a\lambda, b\lambda)$, then we can assume the sum divided by $p$ is still a gamma distribution. Specifically,

$$\lambda_l = \lambda, \quad \lambda \sim \text{Ga}(a\lambda, pb\lambda),$$

where $l = 1, \ldots, p$, and $p$ is the total number of variables. The exponent term in the Gaussian kernel Equation (18) reduced to $-\lambda \sum_{l=1}^{p} (x_l - \chi_l)^2$, which is also the most commonly used kernel in SVM.

When all explanatory variables contain the same amount of information on the response variable, or their difference cannot be detected in a small data set, we use the equal weights prior structure. For example, the ionosphere study (Newman et al., 1998) in Section 4 falls into this category.

2.3.2 BARK with different weights

However, for most problems, given that all explanatory variable are relevant, it is common to believe that they have different effects on the response variable. This translate to different variables contribute to the regression differently through a different kernel scale parameter in BARK. Suppose the prior distribution for the sum of all $\lambda_l$s is $\text{Ga}(a\lambda, b\lambda)$. Notice that gamma distribution is infinitely divisible, we can splits the total gamma mass equally into individual scale parameters $\lambda_l$. Specifically,

$$\lambda_l \overset{\text{iid}}{\sim} \text{Ga}(a\lambda/p, b\lambda),$$

where $l = 1, \ldots, p$, and $p$ is the total number of variables. The independent gamma prior distributions guarantees that all kernel scale parameters are different.

When we believe that all explanatory variables are relevant to the response variable, and there is enough evidence in the data to detect the different contributions in different variables, BARK with different weights are appropriate for the data analysis. The posterior Bayesian inference would shrink the scale parameters for the variables with little effects to values near zero, and features can be selected through the soft shrinkage. For example, the Boston housing data set (Harrison and Rubinfeld, 1978) in Section 4 falls into this category.
2.3.3 BARK with selection and equal weights

Sometimes, hard shrinkage is preferred, \textit{i.e.}, the explanatory variable is either selected and having a reasonable contribution to the response variable, or not selected thus have no contribution to the response variable. This can be achieved by using a prior distribution that is a mixture of a point mass at 0 and a continuous distribution, for example, see George and McCulloch (1997).

In BARK, introduce an indicator vector $\delta \in \{0, 1\}^p$ for the scale parameter $\lambda$ in the kernel function. Typically, we use a Bernoulli prior distribution for each indicators. The use of a hierarchical prior increases the flexibility of the prior distribution and reduces the dependency of the posterior distribution on the prior assumptions. Therefore, making the inclusion probability $p_\lambda$ random is more desirable than fixing it, for example, see Clyde and George (2004); Nott and Kohn (2005). Specifically, the prior distribution for the kernel scale parameters are

$$
\lambda_l = \begin{cases} 
\lambda_\delta^* & \text{if } \delta_l = 1 \\
0 & \text{if } \delta_l = 0 
\end{cases}, \quad \{\delta_l\}_{1 \leq l \leq p} \sim \text{Bi}(1, p_\lambda), \quad p_\lambda \sim \text{Be}(a_p, b_p). \tag{19}
$$

where $\lambda_\delta^*$ is positive for all $l$ with $\delta_l = 1$.

If we believe that the variables that are related to the response are equally important, or if we do not have enough evidence in the data to discriminant the different influence for variables related to the response, we can set all non-zero kernel scale parameters to be the same.

Specifically, on top of Equation (19), let

$$
\lambda^*_l = \lambda \sim \text{Ga}(a_\lambda, b_\lambda),
$$

for all $l$ with $\delta_l = 1$, where $d = \sum_{l=1}^p \delta_l$ is the number of 1s in the indicator vector $\delta$, or the number of non-zero kernel scale parameters. Given $\delta$, the sum of all kernel scale parameters

$$
\sum_{l=1}^p \lambda_l = \sum_{l: \delta_l > 0} \lambda^*_l = d \lambda
$$

has a gamma distribution with shape $a_\lambda$ and scale $b_\lambda$.

In Section 4 we use the Circle simulation studies to illustrate how the prior distributions work. The simulation studies are cooked in a way that the data is generated from models with equal weights signals and some pure noise. It shows that BARK with selection and equal weights can effectively select those signal dimensions, and drop the noise dimensions out.

When not all variables are relevant, and the sample size is not big enough to catch up the different effects of the signal variables, BARK with selection and equal weights can be used for both regression and classification problems. For example, the body fat data set (Johnson, 1995) in Section 4 falls into this category.

2.3.4 BARK with selection and different weights

Similarly, allowing the non-zero scale parameters to be different yields the fourth setting. Specifically, on top of Equation (19), let

$$
\lambda^*_l \sim \text{Ga}(a_\lambda/d, b_\lambda), \tag{20}
$$
for all $l$ with $\delta_l = 1$, where $d$ is the number of non-zero kernel scale parameters. Again, it induces the same Gamma prior distribution with shape $a_\lambda$ and scale $b_\lambda$ for the sum of all kernel scale parameters.

This is the most flexible setting which contains both hard shrinkage via mixture prior distribution with point mass at zero, and soft shrinkage by allowing different non-zero kernel scale parameters. However, this is also the most demanding for the data set. In other words, it requires more samples if the number of explanatory variables are kept fixed in order to both filter out the irrelevant dimensions, and detect the differences within the signal dimensions.

2.4 Elicitation

There are three parameters in the SoS prior specification, $\{\alpha, \gamma, \epsilon\}$. In the continuous approximation, the stable index $\alpha$ serves as the degree of freedom in the student $t$ prior distribution for the regression coefficient $\beta$. In particular, $\alpha = 1$ corresponds the Cauchy random field, which induces Cauchy prior distribution on the regression coefficient $\beta$. Our experience suggest that $\alpha = 1$ is a pretty good default choice, and it works well for both the simulation studies and the real data analysis that we have tried.

Set $\alpha = 1$, we detail the elicitation for $\gamma$ and $\epsilon$ as follows. The approximation threshold $\epsilon$ serves as the scale parameter in the Cauchy prior distribution for the regression coefficient $\beta$, and both $\epsilon$ and the intensity parameter $\gamma$ determines the number of kernels $J$ in BARK. With the continuous approximation, $J$ has a Poisson prior distribution, with mean $\gamma/\epsilon$. In addition, it is desirable to control the level of approximation through the $L_2$ discrepancy Equation (15). In particular, for kernel function $k(x, \chi)$ with a discrete uniform prior for $\chi$ on $n + 1$ locations, the $L_2$ discrepancy

$$E[\mathcal{L}[k(x, \cdot)] - \mathcal{L}_c[k(x, \cdot)]^2 = \gamma \epsilon \|k(x, \cdot)\|_2^2 = \frac{\gamma \epsilon}{n+1} \sum_{i=0}^{n} k(x, \chi_i)^2$$

With a smaller $\epsilon$, the $L_2$ discrepancy is smaller, but we need to use more kernels to make up the smaller regression coefficient associated with each kernel function. We recommend eliciting $\gamma$ and $\epsilon$ through observable quantities the expected number of kernels $\gamma/\epsilon$ and the approximation factor $\gamma \epsilon$. In practice, we found that $\gamma = 5$ and $\epsilon = 0.5$ works well for various simulated data sets and real data sets. Although the approximation is very crude under this setting, the nonparametric model is flexible enough to obtain good predictions.

The spread of the kernel function Equation (18) is controlled by the scale parameters $\lambda_l$’s. Since the variables $X_l$ are standardized to have mean 0 and variance 1 before the analysis, if the kernel center $\chi$ is also standardized, and is independent of $X_l$, the square differences $(x_l - \chi_l)^2$ are independent, with mean 2 and variance about 8 with normal approximations for $X$ and $\chi$. When $\lambda_l$’s are very close to zero, the kernel function is similar to a point mass at 1; when $\lambda_l$’s are very large, the kernel function is similar to a point mass at 0. These cases need to be avoid, because we do not want the kernel behave like the trivial intercept kernel $k(x, x_0)$. Four different prior distributions are specified in Section 2.3, and we recommend using fixed hyperparameters $a_\lambda = b_\lambda = 1$, because they lead to well behaved kernel functions. To be more specific, let $S = -\sum_{l=1}^{p} \lambda_l(x_l - \chi_l)^2$. Assuming that
\[(x_i - \lambda_i)^2/2 \sim \text{Ga}(1/2, 1/2),\] the mean and variance of \(S\) is \(\frac{2a_\lambda}{b_\lambda}\) and \(\frac{4a_\lambda^2}{b_\lambda^2} \left(\frac{1}{a_\lambda} + \frac{2}{d} + \frac{2}{da_\lambda}\right)\) respectively when non-zero \(\lambda_i\)'s are equal, or \(-\frac{2a_\lambda}{b_\lambda}\) and \(\frac{4a_\lambda^2}{b_\lambda^2} \left(\frac{3}{a_\lambda} + \frac{2}{d}\right)\) respectively when non-zero \(\lambda_i\)'s are different, where \(d\) is the number of non-zero \(\lambda_i\)'s. In particular, when \(a_\lambda = b_\lambda = 1\), those numbers are \(-2\) and \(4\) in BARK with equal weights, or \(-2\) and \(12\) in BARK with different weights. Although it is difficult to obtain the exact distribution for \(e^S\), we verified that the inter-quartile range of the kernel function with \(a_\lambda = b_\lambda = 1\) is greater than \(0.5\) for all integer \(d\) with simulations. In other words, half of the mass in the kernel function will span at least length \(0.5\) out of all possible values in \((0,1]\) The typical kernel value \(E(K)\) is in \([0.3,0.5]\), which can be used to elicit the intensity parameter \(\gamma\).

In BARK with selection, the probability that each \(\lambda_i\) is non-zero has a Beta prior distribution, \(\text{Beta}(c,d)\) with similar weights. Although it is difficult to obtain the exact distribution for \(e^S\), we verified that the inter-quartile range of the kernel function with \(a_\lambda = b_\lambda = 1\) is greater than \(0.5\) for all integer \(d\) with simulations. In other words, half of the mass in the kernel function will span at least length \(0.5\) out of all possible values in \((0,1]\) The typical kernel value \(E(K)\) is in \([0.3,0.5]\), which can be used to elicit the intensity parameter \(\gamma\).

In BARK with selection, the probability that each \(\lambda_i\) is non-zero has a Beta prior distribution in Equation (19). This generates a prior distribution for the number of non-zero signal parameters \(d\) that corresponds to the Binomial-Beta distribution (see Bernardo and Smith, 1994, p. 117), with probability mass function

\[
P(d = k) = \binom{p}{k} \frac{\Gamma(a_p + b_p)\Gamma(a_p + k)\Gamma(b_p + p - k)}{\Gamma(a_p)\Gamma(b_p)\Gamma(a_p + b_p + p)}
\]

We recommend using the uniform hyper-prior distribution, \(\text{Ga}(c,d)\) for the overall precision \(\phi\). Since \(\phi\) is always in the model, we can set \(c = d = 0\), which reduces to the non-informative prior with \(\pi(\phi) \propto 1/\phi\). This is an improper prior distribution, but yields proper posterior distribution for the model Equation (7).

### 2.5 Inference

From the independence assumption for \(y_i\), the likelihood for the training data set can be written as

\[
p(y \mid \phi, J, \beta, \chi, \lambda) = \frac{\phi^{n/2}}{(2\pi)^{n/2}} \exp \left\{ \frac{\phi}{2} \|y - K\beta\|^2 \right\}
\]

where \(y = \{y_1, \ldots, y_n\}^T\) and \(K\) is the \(n \times J\) kernel matrix, with \(K_{ij} = K(x_i, \chi_j)\).

Having defined the prior distributions and calculated the likelihood, the Bayesian inference relies on sampling the parameters from the posterior distribution

\[
p(\phi, J, \beta, \chi, \lambda \mid y) = \frac{p(y \mid \phi, J, \beta, \chi, \lambda)\pi(\phi, J, \beta, \chi, \lambda \mid y)}{p(y)}
\]

where \(p(y)\) is the marginal likelihood of \(y\) which integrates out all parameters. Given a new observation at \(x^*\), the predictive distribution for \(y^*\) is

\[
p(y^* \mid y) = \int p(y^* \mid x^*, \phi, J, \beta, \chi, \lambda)p(\phi, J, \beta, \chi, \lambda \mid y) d\phi dJ d\beta d\chi d\lambda
\] (21)
If we have a random sample from the posterior distribution, simulate \( y \) conditional on those sampled parameters, which is the predictive distribution for \( y^* \). In practice, we use the MCMC draws as samples drawn from the posterior distribution.

With the truncation on \( \beta \) approximation to the SoS random field, the regression coefficients have independent symmetric Pareto prior distributions with density

\[
f(\beta) = \frac{\alpha \epsilon^\alpha}{2} |\beta|^{-\alpha-1} 1_{\{|\beta| > \epsilon\}}
\]

There is no conjugate update for the regression coefficient, and sampling from its conditional posterior distribution relies on Metropolis-Hasting updates. Because \( \beta \) is highly correlated with the unknown regression mean, the Markov chain converges very slowly.

Similarly, it is not so convenient to work with \( t_\alpha \) prior distributions Equation (17) directly on the regression coefficients directly in the continuous approximation. However, we can represent \( t_\alpha \) as a mixture of normal distributions, improving the mixing of the Markov chain by integrating out the regression coefficients and making inference on the precision parameters. In particular, when \( \alpha = 1 \), we can collapse the regression coefficient with the same kernel locations. To be more specific, decompose the Cauchy prior distributions to

\[
\tilde{\beta}_i \overset{\text{ind}}{\sim} \text{No}(0, n_i^2 \tilde{\varphi}_i^{-1}), \quad \tilde{\varphi}_i \overset{\text{ind}}{\sim} \text{Ga}\left(\frac{1}{2}, \frac{\epsilon^2}{2}\right), \quad i \in \{i \mid n_i > 0\}.
\]

Conditional on the number of kernels \( J \) and kernel locations \( \{\mathbf{x}_1, \ldots, \mathbf{x}_J\} \), we can integrate the regression coefficient \( \beta \) and use \( \varphi \) to replace the role of \( \beta \) in the likelihood function. In the collapsed representation, denote the index of the non-zero elements in \( n \) is \( i = (i_1, \ldots, i_m) \), i.e., \( n_{ij} > 0 \) for \( j = 1, \ldots, m \). Let \( \beta^* \) and \( \varphi^* \) be the length-\( m \) sub-vector of \( \beta \) and \( \varphi \), where \( \beta_j^* = \beta_{ij}, \varphi_j^* = \varphi_{ij} \). The \( n \times (n + 1) \) kernel matrix \( \tilde{K} \) is defined by \( \tilde{K}_{jk} = k(\mathbf{x}_j, \mathbf{x}_k) \), where \( 1 \leq j \leq n, 0 \leq k \leq n \). Denote by \( K^* \) the \( n \times m \) sub-matrix of \( \tilde{K} \), where \( K_{ik}^* = \tilde{K}_{ij} \).

After integrating out \( \beta^* \), the likelihood is

\[
p(y \mid n, \varphi^*, \lambda, \phi) = \int p(y \mid n, \beta^*, \varphi^*, \lambda, \phi)p(\beta^* \mid \varphi^*) \, d\beta^*
\]

\[
= \phi^{n/2} \prod_{j=1}^{m} \left( \frac{\varphi_j^* / n_j^2}{(2\pi)^{n/2} |\Sigma^*|^{-1/2}} \right) \exp \left\{ -\frac{1}{2} \left( \phi \|y - K^* \mu^*\|^2 + \sum_{j=1}^{m} \frac{\varphi_j^*}{n_j^2} \mu_j^* 2 \right) \right\},
\]

where

\[
\Sigma^* = \left( \phi K^* T K^* + \text{diag} \left( \frac{\varphi_1^*}{n_1^2}, \ldots, \frac{\varphi_m^*}{n_m^2} \right) \right)^{-1}, \quad \mu^* = \phi \Sigma^* K^* T y.
\] (22)

Denote \( \theta = (n, \varphi^*, \lambda, \phi) \), after integrating out \( \beta \), instead of sampling from full joint posterior distribution, we only need to sample from \( p(\theta \mid y) \). Conditional on \( \theta \), the posterior distribution for \( \beta^* \) is \( \text{No}(\mu^*, \Sigma^*) \), where \( \mu^* \) and \( \Sigma^* \) is defined in Equation (22). Given a new observation \( x \), suppose \( \{\theta^{(m)}\}_{m=1}^{M} \) are samples of \( \theta \) from the posterior MCMC, then
\[ \frac{1}{M} \sum_{m=1}^{M} f(x \mid \theta^{(m)}, \mu^{(m)}) \] is a point estimator for the prediction \( y \). This estimator has a smaller variance than \( \frac{1}{M} \sum_{m=1}^{M} f(x \mid \theta^{(m)}, \beta^{(m)}) \) with \( \beta^{(m)} \sim \text{No}(\mu^{(m)}, \Sigma^{(m)}) \) due to Rao-Blackwellization.

Because the dimension of \( \theta \) is not fixed, we use Reversible Jump Monte Carlo Markov chain (RJ-MCMC) algorithm to sample from the posterior distribution, details see appendix B. By integrating out \( \beta \), we reduced the correlation of \( \theta \) and the unknown regression mean function. By sacrificing the conjugacy for \( \phi \) and \( \beta \) in the Gibbs algorithm, we benefit from the weak correlation, which results a better mixed Monte Carlo Markov chain.

3. Bayesian Additive Classification Kernels

We call the classification counterpart of BARK as Bayesian Additive Classification Kernels (BACK), which augment latent random variables to represent the discrete class labels, as shown in Albert and Chib (1993). In this paper, we focus on binary classification, where the response variable \( y \in \{0, 1\} \). With the Probit link function,

\[ P(y_i = 1 \mid x_i) = \Phi(f(x_i)), \]

where \( \Phi(\cdot) \) is the cumulative distribution function for standard normal distribution, we can decompose the model into

\[ y_i = \mathbf{1}_{\{z_i > 0\}}, \quad z_i \overset{iid}{\sim} \text{No}(f(x_i), 1). \]

Conditional on \( z \), this is exactly the BARK we described in the previous section, except that \( \phi \) is fixed at 1 in the Probit model. Another difference in this specification is that we need to obtain the values of \( f(x) \) in order to update \( z \). Previously, we integrated out \( \beta \) in the regression for better mixing Markov chains, but now we need to obtain those regression coefficients to calculate the mean function \( f(x) \) explicitly. Notice that the conditional posterior distribution for \( \beta \) is Normal with mean \( \mu \) and variance \( \Sigma \) as described in Equation (22), the Gibbs sampling is straightforward. After obtaining \( f(x) \), we can sample \( z \) from its conditional distribution. If \( y_i = 1 \), the conditional distribution for \( z_i \) is \( \text{No}(f(x_i), 1) \) truncated above zero; if \( y_i = 0 \), the conditional distribution for \( z_i \) is \( \text{No}(f(x_i), 1) \) truncated below zero.

As a result, given a new observation \( x \), we cannot use the Rao-Blackwellization trick as in the regression case. Instead, the prediction for \( y \) is obtained by the sign of the auxiliary variable \( z = \frac{1}{M} \sum_{m=1}^{M} z^{(m)} \), where \( z^{(m)} = f(x \mid \theta^{(k)}, \beta^{(k)}) \), and \( (\theta^{(k)}, \beta^{(k)}) \) are posterior samples from the Markov chain.

4. Simulation Studies and Examples

We present the summaries of the performance of BARK for both regression and classification problems on some example data sets, comparing results with support vector machine (SVM) and Bayesian adaptive regression tree (BART) for illustrative purposes. Before doing the analysis, we standardized all covariates to have mean 0 and standard deviation 1. For all studies, the hyperparameters are chosen to be \( \alpha = 1, \epsilon = 0.5, \gamma = 5, a_\lambda = b_\lambda = a_p = b_p = 1 \). We discard the initial 2,000,000 iterations for burn in, and keep the chain running for
additional 2,000,000 iterations. For practical reasons, we only keep 4000 samples (one out of every 500) in the Markov chain in the posterior inference. For each simulation study, we use 1000 additional data points to evaluate the predictive performance; for each real data set, we use 5-fold cross-validation to evaluate the predictive performance, and we repeat each experiment 20 times.

4.1 Regression Examples

For regression problems, we calculate predictive mean square error and normalize it with respect to the best method for each run, and then report the average of 20 replicated runs, see Table 1. We demonstrate the performance of our model by three simulation studies and three real data sets.

Table 1: Predictive mean square errors in regression problems.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>BARK</th>
<th>SVM</th>
<th>BART</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>equal</td>
<td>diff</td>
<td>select + equal</td>
</tr>
<tr>
<td>Friedman1</td>
<td>7.31</td>
<td>1.22</td>
<td>2.26</td>
</tr>
<tr>
<td>Friedman2</td>
<td>1.99</td>
<td>1.07</td>
<td>1.09</td>
</tr>
<tr>
<td>Friedman3</td>
<td>3.07</td>
<td>1.46</td>
<td>2.30</td>
</tr>
<tr>
<td>Boston Housing</td>
<td>1.44</td>
<td>1.09</td>
<td>1.23</td>
</tr>
<tr>
<td>Body Fat</td>
<td>1.39</td>
<td>1.81</td>
<td>1.01</td>
</tr>
<tr>
<td>Basketball</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
</tr>
</tbody>
</table>

The simulation studies, Friedman 1, 2, and 3, are described in (Friedman, 1991; Breiman, 1996). The Friedman 1 data set uses 10 independent variables uniformly distributed on the interval [0, 1], and the regression mean function only depend on the first five variables,

$$f_1(x) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5.$$  

Standard normal noise is added to the mean function, so that the signal to noise ratio is 3.28.

The Friedman 2 and 3 data sets use four independent variables that are uniformly distributed over the ranges

$$0 \leq x_1 \leq 100, \quad 40\pi \leq x_2 \leq 560\pi, \quad 0 \leq x_3 \leq 1, \quad 1 \leq x_4 \leq 11.$$  

The corresponding regression mean functions are

$$f_2(x) = (x_1^2 + (x_2x_3 - 1/(x_2x_4))^2)^{1/2},$$  

$$f_3(x) = \arctan((x_2x_3 - 1/(x_2x_4))/x_1).$$

Independent Gaussian noise with mean zero and standard deviations 125 and 0.1 are added to the mean function in Friedman 2 and 3 data sets respectively, so that the signal to noise ratio is approximately 3.

In Friedman 1 simulation study, since there are five noise variables, both BART with equal weights and SVM won’t work so well, and they have a relatively higher out-of-sample
mean square errors. However, other BART models with either soft or hard shrinkage can filter out the noise variables, and obtain good out-of-sample predictions. In Friedman 2 and Friedman 3 simulation studies, there are 200 samples with only four variables. Therefore, we have enough data to detect the different effects among the variables, and BARK with different weights performs better than BARK with equal weights. Although BARK lose to BART for Friedman 3 simulation study, the overall performance of BARK is comparable with BART.

The Boston housing data set (Harrison and Rubinfeld, 1978) contains 506 data points with 13 covariates, and the goal is to predict the median home value. The data set is originally proposed to address how does the environmental conditions affect the housing price. It is also a well studies data set for variable selection in statistics literature. For example, Breiman and Friedman (1985) discovered that RM, TAX, PTRATIO and LSTAT were the four most important variables using ACE transformations, and Smith and Kohn (1996) argues that NOX, RM, DIS, TAX and LSTAT were most important using Bayesian variable selection. Although BART beat the BARK model for this data set, BARK with either soft or hard shrinkage beats SVM, which has no feature selection property. Figure 1 shows the box plot for the those scale parameters in the Boston housing data set in model BARK with different weights. A larger value on $\lambda$ corresponds more influence on the regression mean through the kernel function. As we can see, kernel scale parameters correspond to NOX, RAD and LSTAT are significantly bigger than others, hence these variables are crucial in the prediction of the median housing price. On the other hand, we see $\lambda$s on ZN, CHAS and B are tiny, hence we conclude that these three variables does not effect the housing price much.

The body fat data set (Johnson, 1995) lists estimates of percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. The goal is to use 14 relevant covariates to predict the body fat percentage. The cross-validation results from Table 1 suggests that model BARK with selection and equal weights makes the best prediction. In fact, the body fat percentage can be well predicted by just two variables, the density determined from underwater weighing and the wrist circumference, as shown in the box plots of the scale parameters for each variable in Figure 2.

The basketball data set (Simonoff, 1996, p. 276) contains demographic and box-score statistics for 96 players. The goal is to predict the points scored per minute played from assist credited per minute played, height, minute played per game and age. Table 1 shows that the BARK models, SVM and BART are comparable in terms of out-of-sample prediction mean square errors.

### 4.2 Classification Examples

For classification problems, we calculate the predictive misclassification rate, and report the average of the 20 replicated runs in table 2. We demonstrate the performance of our model with three simulated studies and three real data sets from the UCI Machine Learning Repository.

The simulation studies are called Circle 2, 5, 20, which have 2, 5, 20 variables respectively. All variables are generated from a uniform distribution in $[-1, 1]$, but the class label
only depend on the first two variable, \( y = 1 \{ x_1^2 + x_2^2 \leq 2/\pi \} \). Under this formulation, both class have roughly the same number of sample points.

Table 2: Predictive mis-classification rate in classification problems.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>BARK</th>
<th>SVM</th>
<th>BART</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>equal</td>
<td>diff</td>
<td>select + equal</td>
</tr>
<tr>
<td>Circle 2</td>
<td>1.93%</td>
<td>4.91%</td>
<td>1.88%</td>
</tr>
<tr>
<td>Circle 5</td>
<td>13.50%</td>
<td>4.70%</td>
<td>1.47%</td>
</tr>
<tr>
<td>Circle 20</td>
<td>49.16%</td>
<td>4.84%</td>
<td>2.09%</td>
</tr>
<tr>
<td>Bank</td>
<td>1.05%</td>
<td>1.25%</td>
<td>0.55%</td>
</tr>
<tr>
<td>WDBC</td>
<td>2.70%</td>
<td>4.02%</td>
<td>2.49%</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>5.33%</td>
<td>8.59%</td>
<td>5.78%</td>
</tr>
</tbody>
</table>
Figure 2: Box plots for the kernel scale parameters in body fat data set in BARK with selection and equal weights.

As we can see from Table 2., SVM suffers greatly from the increasing noisy dimensions, and so does BARK with equal weights. Using a common scale parameter for all covariates in the kernel function won’t work when there are a lot of noisy variables. On the other hand, other BARK models with feature selection property does not suffer from this problem as the number of noisy dimensions increases. The prior distribution enables the model to automatically shrink the contributions of the noisy dimensions to zero or negligible values, thus focusing the regression only on the first two signal dimensions. Under the simulation setting, it is clear to see that BARK with selection and equal weights is the most efficient. It is not surprising, because the simulated data are generated with equal signal dimensions, where as the remaining noisy dimensions do not contributed to the classification at all. The posterior probability that $\lambda_l > 0$ are 1, 0.996 for the first two dimensions, and near zero for the rest 18 dimensions, which confirm that BARK with selection and equal weights indeed focused on the first two dimensions in making the classification. Other BARK models with feature selection property does not take advantage of “knowing the variable structure ahead of time and building it into the prior distribution”, they are less efficient than BAKR with selection and equal weights, however, they still beat BART in the out-of-sample prediction.
The Swiss bank notes data (Flury and Riedwyl, 1988, p. 5) contains 100 genuine notes \((y = 0)\) and 100 counterfeit notes \((y = 1)\). There are six predictors, each giving a different aspect of the size of the note: the bottom edge length, the diagonal length, the left edge length, the center length, the right edge length and the top edge length. The task is to identify counterfeit notes from these six features. Table 2 suggests that BARK with selection and equal weights has a very good out-of-sample prediction. However, the posterior probability that \(\lambda_i > 0\) are for the six predictors are 0.14, 0.252, 0.302, 0.776, 0.278, 0.962 respectively, and the box plots for those kernel scale parameters in shown in Figure 3. This is very different from Circle 20 simulation study or the Body Fat data set, where BARK with selection and equal weights suggests that the posterior model only contains two variable, while BARK with selection and equal weights for Swiss Bank data set suggest that the posterior model is a complicated mixture of all six variables. A closer look at the prediction process in BARK reveals that the prediction is actually based on an average estimator from lots of posterior models. Although each posterior model is a BARK contains a subset of variables with equal weights, the average of all those models can be much more flexible. For this particular data set, parametric models with direct selection on the original six variables is not ideal, but BARK can make very good predictions by jumping among different “variable selection models” in the posterior sampling. In fact, Cook and Lee (1999) suggest to make classification based on two linear combinations of the original six variables.

The Wisconsin diagnostic breast cancer (WDBC) data set (Wolberg et al., 1995) contains 357 benign \((y = 0)\) samples and 212 malignant samples \((y = 1)\) with 30 real-valued geometric features for the cell nucleus. The task is to diagnose cancer from these geometric features. Table 2 suggests that BARK with equal weights have the same performance as that of SVM, which also use Gaussian kernel with a common scale parameter. Because there are 30 different explanatory variables, 506 samples are not big enough to discover different influences of important variables on detecting cancer. Therefore, BARK with different weights, or with selection and different weights do not perform as good as BARK with selection and equal weights.

The original ionosphere data set (Newman et al., 1998) contains 351 observations with 34 measures signals on different frequency domains. Because the second covariate is zero for all observations, we exclude it in the analysis, resulting only 33 effective covariates. The goal is to detect whether there is evidence of some type of structure in the ionosphere. Table 2 shows that BARK with equal weights has the best performance among different BARK models, which is also comparable to SVM and BART. This actually suggest that the signals from all different frequency domains contribute to the structure in the ionosphere, and their contributions are the same.

5. Discussion
In this paper, we have developed a fully Bayesian kernel method, for both nonparametric regression and classification. The model is based on a linear expansion of kernel functions, which combines the interactive effects through addition. The unknown mean function is formulated as a stochastic integral of a kernel function with respect to a random signed measure, which can be approximated by a finite sum of a random number of kernel functions at random locations. The kernel scale parameters are covariate specific and thus adapt to
the features of the data. The RJ-MCMC algorithm developed for fitting the model provides an automatic search mechanism for finding sparse representations of the mean function, and the posterior analysis for the kernel scale parameters provides insights for making feature selection on the original covariates.

The model presented in Section 2 admits a number of extensions. In this paper, we restrict the kernel functions located at the training data points, which reduced the computation when \( n \) is small and \( p \) is large. However, for lower dimensional problems, say \( p = 1 \) or 2, it is more flexible to allow the kernel located at any place on \( \mathbb{R}^p \). One natural extension to the discrete uniform prior distribution on the training sample points is a mixture of continuous distributions centered at the training samples. Then we increased model flexibility, and still focus on exploring the space close the the observed data.

Another restriction in our model is that all kernel functions share the same shape parameter. We can extend the Lévy random field \( \nu \) to \( \mathbb{R} \times X \times \Lambda \), where \( \Lambda \) is the space of the kernel scale parameters. Then the model induced by this prior specification will have kernel functions with different shape parameters. This allows the model to adopt different local features at different regions of the sample space. Notice that the parameter space increases
greatly under this formulation, so it may be necessary to use a sparse prior distribution on \( \Lambda \) such that most of the scale parameters are zero for each kernel.

We only demonstrated our model for binary classification in Section 3, but it is straightforward to extend this model to ordered multi-class case. For \( d \) different classes, introduce \( d + 1 \) cut-off real values \(-\infty = c_0 < c_1 = 0 < c_2 < \cdots < c_{d-1} < c_d = \infty\). Use the same latent normal random variable \( z \), such that \( y \) is in class \( k \) if \( c_{k-1} < z < c_k \). Incorporating the update schemes for \( c_k \), the model described in Section 3 is applicable for the multi-class classification problems.

Acknowledgments

This work was supported in part by National Science Foundation grants DMS–0342172, DMS-0757549, DMS–0635449 and PHY–0941373 and NASA grant NNX09AK60G. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
Appendix A. Proof of Theorem 1

For any \( g \in L_2(\mathcal{X}, \pi(d\chi)) \), rewrite the Poisson construction Equation (12) as

\[
\mathcal{L}[g] = \iiint_{(-1,1) \times (0,1)} \beta g(\chi) N(d\beta, d\chi, du) + \iiint_{(-1,1) \times (0,1)} (\beta - \sin \beta) g(\chi) \nu(d\beta, d\chi, du) + \frac{\|g\|_2^2}{2} g^2(\chi) \pi(d\chi)
\]

(23a)

\[
= \iiint_{(-1,1) \times (0,1)} \beta g(\chi) N(d\beta, d\chi, du) - \iiint_{(-1,1) \times (0,1)} \sin \beta g(\chi) \nu(d\beta, d\chi, du).
\]

(23b)

Because \( \beta - \sin \beta = O(\beta^2) \) when \( \beta \approx 0 \), and \( 1_{\{\beta < 1\}} \beta^2 g(\chi) \) is \( \nu \)-integrable, the second term in Equation (23b) is finite. In addition, \( \beta - \sin \beta \) is an odd function, and \( \nu(d\beta, d\chi) \) is symmetric about zero on the first dimension, so \( \int_{(-1,1) \times (0,1)} (\beta - \sin \beta) g(\chi) \nu(d\beta, d\chi) du = 0 \). Similarly, the fourth terms in Equation (23b) is also zero.

Notice that the difference between Equation (23b) and Equation (13) is

\[
\mathcal{L}[g] - \mathcal{L}_\epsilon[g] = \iiint_{\mathbb{R} \times \mathcal{X} \times (0,1)} 1_{\{1 + \alpha_1^2 \beta^2 - (\alpha + 1)/2 \leq u < 1\}} \beta g(\chi) N(d\beta, d\chi, du) + \frac{\|g\|_2^2}{2} g^2(\chi) \pi(d\chi)
\]

\[
= \iiint_{(-1,1) \times (0,1)} 1_{\{1 + \alpha_1^2 \beta^2 - (\alpha + 1)/2 \leq u < 1\}} \beta g(\chi) N(d\beta, d\chi, du).
\]

(24)

Since \( g \in L_2(\mathcal{X}, \pi(d\chi)) \), \( \|g\|_2^2 = \int_{\mathcal{X}} g^2(\chi) \pi(d\chi) < \infty \). The \( L_2 \) discrepancy of this approximation is

\[
E \left| \mathcal{L}[g] - \mathcal{L}_\epsilon[g] \right|^2 = \iiint_{\mathbb{R} \times \mathcal{X} \times (0,1)} 1_{\{1 + \alpha_1^2 \beta^2 - (\alpha + 1)/2 \leq u < 1\}} \beta^2 g^2(\chi) \nu(d\beta, d\chi, du)
\]

\[
= \|g\|_2^2 \frac{2\gamma(\alpha + 1)}{\pi} \sin \left( \frac{\pi\alpha}{2} \right) \int_0^\infty \left( 1 - \left( 1 + \frac{\alpha_1^2 \beta^2}{\beta^2} \right) \right)^{\frac{1+\alpha}{\beta}} \beta^{1-\alpha} d\beta,
\]

(24)

Set \( \delta = 1/(\epsilon \sqrt{\alpha}) \), then when \( \beta > \delta \), \( 0 < \alpha_1^2 \beta^2 - 1 < 1 \). From the binomial theorem,

\[
\left( 1 + \frac{\alpha_1^2 \beta^2}{\beta^2} \right)^{-(1+\alpha)/2} \geq 1 - \frac{\alpha(1 + \alpha)\epsilon^2}{2\beta^2}.
\]

Therefore,

\[
\int_0^\infty \left( 1 - \left( 1 + \frac{\alpha_1^2 \beta^2}{\beta^2} \right)^{-(1+\alpha)/2} \right) \beta^{1-\alpha} d\beta \leq \int_0^\infty \frac{\alpha(1 + \alpha)\epsilon^2}{2\beta^{1+\alpha}} d\beta = \frac{(1 + \alpha)\epsilon^2}{2\delta^\alpha} = \frac{(1 + \alpha)\alpha^{\alpha/2} \epsilon^{2-\alpha}}{2}
\]

In addition,

\[
\int_0^\delta \left( 1 - \left( 1 + \frac{\alpha_1^2 \beta^2}{\beta^2} \right)^{-(1+\alpha)/2} \right) \beta^{1-\alpha} d\beta \leq \int_0^\delta \beta^{1-\alpha} d\beta = \frac{\delta^{2-\alpha}}{2 - \alpha} = \frac{\alpha(2-\alpha)/\epsilon^{2-\alpha}}{2 - \alpha}.
\]
Combining these two bounds into Equation (24),

\[ E \left| \mathcal{L}[g] - \mathcal{L}_t[g] \right|^2 \leq \|g\|_2^2 \gamma \Gamma(\alpha + 1) \sin \left( \frac{\pi \phi}{2} \right) \left( \frac{(1 + \alpha)\alpha^{\alpha/2}}{2} + \frac{\alpha^{(2-\alpha)/2}}{2 - \alpha} \right) \epsilon^{2-\alpha}. \]

Because \( 0 < \alpha < 2 \), the above term goes to zero as \( \epsilon \) approaches zero. In conclusion, \( \mathcal{L}[g] - \mathcal{L}_t[g] \) converges to 0 in \( L_2 \) for any \( g \in L_2 \). 

**Appendix B. Details on the MCMC for BARK**

We shall use reversible jump Monte Carlo Markov Chain (RJ-MCMC) algorithm (Green, 1995) to implement this trans-dimensional Markov chain.

For regression problems, use the same notations in Section 2.4, the full parameter set \( \{J, \beta, \varphi, \chi, \lambda, \phi\} \) reduced to \( \{n, \beta^*, \varphi^*, \lambda, \phi\} \) in the collapsed representation. Since \( \beta^* \) is integrated out in the likelihood, we only need to sample \( (n, \varphi^*, \lambda, \phi \mid y) \).

1. Update \( (n, \varphi^*) \) using RJ-MCMC algorithm, conditioned on other parameters.

2. Update \( \lambda \) using standard Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), conditioned on other parameters. Pick one of \( \lambda_i \), say \( \lambda_{i(prop)} \) from vector \( \lambda \) at random, and update it via normal random walk on the log scale to \( \lambda_{i(prop)}^{(prop)} \). The acceptance rate is the minimum of 1 and

\[ \frac{p(y \mid \theta^{(prop)}) \pi_\lambda(\lambda_{i(prop)}^{(prop)}) \lambda_{i(prop)}^{-1}}{p(y \mid \theta) \pi_\lambda(\lambda_{i(prop)}^{(prop)} \lambda_{i(prop)}^{-1})}. \]

3. Update \( \phi \) using standard Metropolis-Hastings algorithm, conditioned on other parameters. The prior density for \( \phi \) is proportional to \( \phi^{-1} \), which cancels the proposal density, hence the acceptance rate is the minimum of 1 and

\[ \frac{p(y \mid \theta^*) \phi^*^{-1} \phi^{-1}}{p(y \mid \theta) \phi^{-1} \phi^*^{-1}} = \frac{p(y \mid \theta^*)}{p(y \mid \theta)}. \]

For classification problems,

1. Update \( (n, \varphi^*, \lambda) \) as in the regression case (1-2), conditioned on the latent normal random variable \( z \). Notice that \( \phi \equiv 1 \) in the classification case.

2. Simulate \( \beta^* \sim \text{No}(\mu^*, \Sigma^*) \), where \( \mu^* \) and \( \Sigma^* \) is defined in Equation (22).

3. Simulate \( z \) from its full conditional distribution given \( y, n, \beta^* \), i.e.,

\[ z_i \sim \begin{cases} 1_{\{z \geq 0\}} \text{No}(z; K^* \beta^*, 1), & \text{if } y_i = 1 \\ 1_{\{z < 0\}} \text{No}(z; K^* \beta^*, 1), & \text{if } y_i = 0 \end{cases} \]

Now we detail the RJ-MCMC algorithm. Suppose the current \( \theta \) have \( k \) kernels, i.e., \( k = \sum_{i=0}^n n_i \). Set the probability of taking a birth, death, or update step be \( p_b(k), p_d(k), \) or \( p_u(k) \) respectively, with \( p_b(k) + p_d(k) + p_u(k) = 1 \).
1. **Birth.** First, we need to propose a new kernel. Set the new kernel location $\mathbf{x}_j^*$ uniformly from $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$, say $\mathbf{x}_j^* = \mathbf{x}_{i^*}$. If there is already some kernel located at $\mathbf{x}_j^*$, or $n_{i^*} \neq 0$, update $n_{i^*} = n_{i^*} + 1$, and keep $\varphi^*$ unchanged. If no existing kernel located at $\mathbf{x}_j^*$, or $n_{i^*} \neq 0$, update $n_{i^*} = 1$, propose a new regression coefficient precision from the prior distribution $\varphi^* \sim (\alpha/2, \alpha\epsilon^2/2)$, and add it into the current $\varphi^*$.

In order to calculate the acceptance ratio of the proposal, we need to set up a death scheme. Let’s kill an existing kernel with probability proportional to some fixed power $\delta$ of its regression precision. In other words, the probability of selecting kernel at location $\mathbf{x}_i$ to kill is proportional to $n_i^{\delta}$. Denote by $p^{(kill)}$ the probability to kill the newly proposed kernel from the new parameters.

Notice that the Jacobian is 1 under this proposal, hence the acceptance rate is the minimum of 1 and the product of the conditional posterior density ratio and the proposal density ratio,

$$
\frac{p(y \mid \theta^*) \pi(\theta^*) q(\theta \mid \theta^*)}{p(y \mid \theta) \pi(\theta) q(\theta^* \mid \theta)} = \frac{p(y \mid \theta^*) \nu^+(\alpha, \gamma, \epsilon)p_d(k+1)p^{(kill)}}{p(y \mid \theta)n_{i^*}^*p_0(k)}.
$$

2. **Death.** Reverse the birth step, first select one existing kernel to kill, say kernel located at $\mathbf{x}_{i^*}$. Denote by $p^{(kill)}$ the probability to kill that kernel. The acceptance rate is the minimum and

$$
\frac{p(y \mid \theta^*) \pi(\theta^*) q(\theta \mid \theta^*)}{p(y \mid \theta) \pi(\theta) q(\theta^* \mid \theta)} = \frac{p(y \mid \theta^*) n_{i^*}^*p_0(k - 1)}{p(y \mid \theta) \nu^+(\alpha, \gamma, \epsilon)p_d(k)p^{(kill)}}.
$$

3. **Update.** The update step does not change the number of kernels used in the regression.

(a) **Update $\varphi$.** Keep all kernels centered at previous locations, i.e., keep $\mathbf{n}$ fixed, pick one element from vector $\varphi^*$, and update it via normal random walk on the log scale. Suppose we propose $\varphi_{i^*}^{(prop)}$ to $\varphi_{i^*}^{(\text{prop})}$, then the acceptance rate is the minimum of 1 and

$$
\frac{p(y \mid \theta^*) \text{Ga}(\varphi_{i^*}^{(\text{prop})} \mid \alpha/2, \alpha\epsilon^2/2, \varphi_{i^*}^{(\text{prop})})}{p(y \mid \theta) \text{Ga}(\varphi_{i^*}^* \mid \alpha/2, \alpha\epsilon^2/2, \varphi_{i^*}^*)}.
$$

(b) **Update $(\mathbf{n}, \varphi)$.** We can also keep the number of kernels fixed, by proposing a birth and a death step together. First, choose a location in $\{0,1,n\}$ with probability proportional to $\mathbf{n}$, and subtract 1 from that coordinate, then add 1 to a random location in $\mathbf{n}$. Similar to the birth step, if the newly proposed kernel already exist, keep the old $\varphi^*$, otherwise, propose a new $\varphi^*$ from its prior distribution. The multinomial prior density for $\mathbf{n}$ exactly cancels the proposal density, hence the acceptance rate is the minimum of 1 and

$$
\frac{p(y \mid \theta^*)}{p(y \mid \theta)}.
$$
References


