

Bayesian Additive Regression Kernels

Zhi Ouyang Merlise A. Clyde

Robert L. Wolpert

Department of Statistical Science

Duke University, Durham, NC 27708-0251

June 8, 2008

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, which is constructed by a prior using α -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 parameters. We show that the α -stable prior on the kernel regression coefficients may be approximated by t_α distributions. With a heavy tail prior distribution on the kernel regression coefficients and a finite support on the kernel location parameter, BARK achieves sparse representations. The shape parameters in the kernel functions capture the non-linear interactions of the variables, which can be used for feature selection. A reversible-jump Markov chain Monte Carlo algorithm is developed to make posterior inference on the unknown mean function. For binary classification using a Probit link, we augment the model with latent normal variables, hence the same method for Gaussian noise applies in the classification problem. We illustrate the approach on several simulated and real data sets.

Key words: Bayes; Kernel Regression; Classification; Supervised learning; Feature Selection; symmetric α -stable Lévy random field; Reversible Jump Markov Chain Monte Carlo.

1 Introduction

In supervised learning, we are given a set of observed input vectors $\{\mathbf{x}_i\}_{i=1}^n$ along with the responses $\{y_i\}_{i=1}^n$. Typically, the response variable Y is one dimensional, which could be continuous (as in regression) or discrete (as in classification), while the covariate is multi-dimensional, say $\mathbf{X} = (X_1, \dots, X_p) \in \mathbb{R}^p$. The goal is to learn the unknown relationship between the response variable and the covariate from the training data set, hence we can make accurate predictions of Y for a new observation $\mathbf{X} = \mathbf{x}$.

Regression models are typically used to understand the relationship between the response Y and covariate \mathbf{X} . Denote $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = f(\mathbf{x})$, and one popular candidate for the mean function is constructed by the sum of generating functions,

$$f(\mathbf{x}) = \sum_j g(\mathbf{x}, \boldsymbol{\theta}_j), \quad (1)$$

where $\boldsymbol{\theta}_j$ is the parameter in the j th generating function. For example, Bayesian Additive Regression Trees (BART) (Chipman *et al.*, 2007) uses tree models as the generating functions. Alternatively, kernel functions can be used as the generating too. Specifically, $g(\mathbf{x}, \boldsymbol{\theta}_j) = \beta_j K(\mathbf{x}, \boldsymbol{\chi}_j)$, where β_j and $\boldsymbol{\chi}_j$ are the corresponding regression coefficient and the kernel location parameter for the j th kernel. Both the “sum-of-trees” model and the “sum-of-kernels” model explore the additive effects through the linear combination of the different generating functions, and explore the interactive effects through individual non-linear generating functions.

Kernel methods have been studied for a long time in both machine learning and statistics literature, (Hofmann *et al.*, 2008; Pillai *et al.*, 2007). The goals are obtaining a sparse representation with few kernels in the sum, and getting good values for both the weight $\boldsymbol{\beta}$ and the kernel parameter $\boldsymbol{\chi}$ for prediction. The representation (1) with kernel generating functions includes a variety of popular models. For example, the Support Vector Machine (SVM) (Cristianini and Shawe-Taylor, 2000; Boser *et al.*, 1992), uses n kernels that center at every observed point. It seeks the optimal $\boldsymbol{\beta}$ that minimize the error loss function and model complexity, where the prediction is based on

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^n \beta_j K(\mathbf{x}, \mathbf{x}_j).$$

Relevance Vector Machines (RVM) (Tipping, 2001) also uses n kernels centered at training samples. In the regression case, RVM assumes a Gaussian additive noise,

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \text{No}(0, \phi^{-1}). \quad (2)$$

It maximizes the type-II likelihood under following prior distributions,

$$\beta_j \stackrel{iid}{\sim} \text{No}(0, \varphi_j^{-1}), \quad \varphi_j \stackrel{iid}{\sim} \text{Ga}(a, b), \quad \phi \sim \text{Ga}(c, d). \quad (3)$$

Both SVM and RVM search for the regression coefficient $\boldsymbol{\beta}$ that optimizes the target function. Since most of the β_j s in the solution are zero or effectively zero, they all have sparse representations. However, SVM and RVM only deliver a point estimator, and they do not provide the predictive distribution for future observations. In addition, no fully Bayesian procedure can be applied for RVM under the recommended setting $a = b = c = d = 0$. The improper prior distribution for regression coefficient leads to an improper posterior distribution, which is problematic for making inferences.

A fully Bayesian approach has the advantage of making probabilistic statement for the prediction and model parameters. Chakraborty *et al.* (2004) developed a Bayesian version

of SVM and RVM. In their hierarchical Bayes relevance vector machine, with Gaussian noise model (2), they used proper prior distribution for the full Bayesian analysis.

In this paper, we detail another fully Bayesian framework for supervised learning with mean function (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 (3). When b goes to zero, the posterior distribution for β is improper if J is fixed. One way to overcome this impropriety problem is to specify the prior distributions for β and J jointly. A small b yields a large φ and a small β . Since the mean function $f(\mathbf{x})$ 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. 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 and Wolpert (2007); Clyde *et al.* (2006); Tu *et al.* (2006); Clyde *et al.* (2005), and we shall extend this approach to supervised learning problems in this paper.

We introduce a generalization of independent Cauchy prior distributions for non-parametric regressions, which is called symmetric α -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.*, 2004) 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 α -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 hyper-parameters, 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

Given observed covariate vectors $\{\mathbf{x}_i\}_{i=1}^n$ in \mathbb{R}^p and the response $\{y\}_{i=1}^n$, assuming independent additive Gaussian noise, BARK is formulated by

$$y_i = \sum_j \beta_j K(\mathbf{x}, \boldsymbol{\chi}_j) + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \text{No}(0, \phi^{-1}). \quad (4)$$

Notice that the mean function as a weighted sum of kernel functions can be also represented as the integral of the kernel function with respect to a signed Borel measure,

$$f(\mathbf{x}) = \sum_j \beta_j K(\mathbf{x}, \boldsymbol{\chi}_j) = \iint_{\mathbb{R} \times \mathbb{X}} K(\mathbf{x}, \boldsymbol{\chi}) \mathcal{L}(d\boldsymbol{\chi}), \quad (5)$$

where $\mathcal{L}(d\boldsymbol{\chi}) = \sum_j \beta_j \delta_{\boldsymbol{\chi}_j}(\boldsymbol{\chi})$ is the signed Borel measure, which puts mass β_j at location $\boldsymbol{\chi}_j$. A random measure \mathcal{L} induces a linear mapping $g \mapsto \mathcal{L}[g]$ from L_2 functions g to random variables $\mathcal{L}[g] = \int_{\mathbb{X}} g(\boldsymbol{\chi}) \mathcal{L}(d\boldsymbol{\chi})$; such a mapping is called a random field. In particular, for any \mathbf{x} , bounded kernel function $K(\mathbf{x}, \cdot)$ is L_2 integrable on \mathbb{X} with respect to probability measure $\pi_{\mathbf{x}}(d\boldsymbol{\chi})$. Therefore, specifying a prior distribution for the unknown mean function $f(\cdot)$ is equivalent to specifying a prior distribution for the random measure $\mathcal{L}(d\boldsymbol{\chi})$ with a random field.

2.1 Symmetric α -stable Lévy Random Fields

Lévy random field \mathcal{L} is a particular choice for the prior distribution on the random measure $\mathcal{L}(d\boldsymbol{\chi})$. For example, given a finite positive measure $\nu(d\beta, d\boldsymbol{\chi})$ on $\mathbb{R} \times \mathbb{X}$ with mass $\nu^+ = \nu(\mathbb{R} \times \mathbb{X}) = \iint_{\mathbb{R} \times \mathbb{X}} \nu(d\beta, d\boldsymbol{\chi}) < \infty$, $f(\mathbf{x})$ from equation (5) can be evaluated by drawing

$$J \sim \text{Po}(\nu^+), \quad \{(\beta_j, \boldsymbol{\chi}_j)\}_{1 \leq j \leq J} \mid J \stackrel{iid}{\sim} \nu(d\beta, d\boldsymbol{\chi})/\nu^+,$$

where $\nu(d\beta, d\boldsymbol{\chi})$ is called the Lévy measure for the Lévy random field \mathcal{L} . Generally, the Lévy measure do not need to be finite (more details on the general Lévy measure see [Cont and Tankov, 2004](#), pp. 457-459).

The symmetric α -stable ($S\alpha S$) Lévy random field is the limiting case for the prior specification (3) as b goes to zero and the number of kernels n goes to infinity. Denote by the $S\alpha S$ Lévy measure

$$\nu(d\beta, d\boldsymbol{\chi}) = \gamma c_{\alpha} |\beta|^{-1-\alpha} d\beta \pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi}), \quad (6)$$

where $c_\alpha = (\alpha/\pi)\Gamma(\alpha)\sin(\pi\alpha/2)$, and $\pi_{\mathbf{x}}(d\mathbf{X})$ is a probability measure on \mathbb{X} . Here $0 < \alpha < 2$ is called the stable index, and $\gamma > 0$ is called the intensity parameter. It induces a Lévy random measure that maps disjoint Borel sets $A_j \in \mathbb{X}$ to independent infinite divisible stable random variables $\mathcal{L}(A_j) \sim \text{St}(\alpha, 0, \gamma\pi(A_j), 0)$ (see [Samorodnitsky and Taqqu, 1994](#), pp. adding page number).

In practice, Lévy random fields can be constructed from Poisson random measures, which can be further used in making posterior Bayesian inference, see the appendix [A](#) or [Tu et al. \(2006\)](#) for details. When the stable index α is equal or greater than 1 in the $S\alpha S$ Lévy random field, compensator functions are required. Luckily, notice that the $S\alpha S$ Lévy measure [\(6\)](#) is symmetric about 0 on β , the effects of any odd compensator function cancel out (see [Tu et al., 2006](#); [Sato, 1999](#), pp. 38).

Since the symmetric $S\alpha S$ Lévy random measure is not finite, approximations are required to generate samples from the random variable $f(x)$ in [\(5\)](#). One common approach is to truncate the mass β with respect to a given threshold $\epsilon > 0$. The Lévy measure is approximated by

$$\nu_\epsilon^T(d\beta, d\mathbf{X}) = \gamma c_\alpha |\beta|^{-1-\alpha} \mathbf{1}_{\{|\beta| > \epsilon\}}(\beta) d\beta \pi_{\mathbf{X}}(d\mathbf{X}),$$

which has a finite mass

$$\nu_\epsilon^{T+}(\alpha, \gamma, \epsilon) = \frac{2\gamma\Gamma(\alpha)}{\pi\epsilon^\alpha} \sin\left(\frac{\alpha\pi}{2}\right).$$

In particular, for the Cauchy random field, $\alpha = 1$ and $\nu_\epsilon^+ = (2\gamma)/(\pi\epsilon)$.

The truncation approximation yields a finite Lévy measure, which induces a joint prior distribution for the number of kernels J , regression coefficient $\boldsymbol{\beta}$ and kernel locations $\boldsymbol{\chi}$ as follows,

$$J \sim \text{Po}(\nu_\epsilon^{T+}), \quad \{\beta_j\}_{1 \leq j \leq J} \stackrel{iid}{\sim} \frac{\alpha\epsilon^\alpha}{2} |\beta|^{-\alpha-1} \mathbf{1}_{|\beta| > \epsilon} d\beta, \quad \{\boldsymbol{\chi}_j\}_{1 \leq j \leq J} \stackrel{iid}{\sim} \pi(\boldsymbol{\chi}), \quad (7)$$

where the prior distribution for the regression coefficients β 's are called two-sided Pareto distributions. The approximated Lévy random Field \mathcal{L}_ϵ^T maps function g to $\mathcal{L}_\epsilon^T[g]$. [Tu et al. \(2006\)](#) has shown that $\mathcal{L}_\epsilon^T[g]$ converges to $\mathcal{L}[g]$ in L_2 , and the expected squared discrepancy of the truncation approximation is finite:

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

or $(2\gamma\epsilon/\pi)\|K(\mathbf{x}, \cdot)\|_2^2$ for the Cauchy case with $g(\boldsymbol{\chi}) = K(\mathbf{x}, \boldsymbol{\chi})$.

Although truncating facilitates the Bayesian inference with $S\alpha S$ Lévy random field, the mixing of the Markov chain in practice is not satisfactory due to the nature of hard cut-off in truncating β . Alternatively, we approximate the $S\alpha S$ Lévy random field continuously by the following Lévy measure

$$\nu_\epsilon^C(d\beta, d\mathbf{X}) = \gamma c_\alpha (\beta^2 + \alpha\epsilon^2)^{-(\alpha+1)/2} d\beta \pi_{\mathbf{X}}(d\mathbf{X}),$$

which has a finite mass

$$\nu_\epsilon^{C^+}(\alpha, \gamma, \epsilon) = \frac{\gamma \alpha^{1-\alpha/2}}{2^{1-\alpha} \epsilon^\alpha} \frac{\Gamma(\alpha/2)}{\Gamma(1-\alpha/2)}. \quad (8)$$

In particular, for the Cauchy random field, $\alpha = 1$ and $\nu_\epsilon^+ = \gamma/\epsilon$.

The continuous approximation also yields a finite Lévy measure, which induces a different joint prior distribution for the number of kernels J , regression coefficient β and kernel locations \mathbf{x} ,

$$J \sim \text{Po}(\nu_\epsilon^{C^+}(\alpha, \gamma, \epsilon)), \quad \{\beta_j\}_{1 \leq j \leq J} \stackrel{iid}{\sim} t(\alpha, 0, \epsilon^2), \quad \{\mathbf{x}_j\}_{1 \leq j \leq J} \stackrel{iid}{\sim} \pi(\mathbf{x}), \quad (9)$$

where the density function for a student t distribution $t(\alpha, 0, \epsilon^2)$ is

$$\pi_\epsilon(d\beta) = \frac{\Gamma((\alpha+1)/2) / \Gamma(\alpha/2)}{(\alpha \epsilon^2 \pi)^{1/2}} \left(1 + \frac{\beta^2}{\alpha \epsilon^2}\right)^{-\frac{\alpha+1}{2}} d\beta,$$

The stable index α automatically becomes the degree of freedom in the t distribution in the approximation. As ϵ goes to zero, the random variable $f(x)$ in (5) constructed from the approximated Lévy random field converges to the one without approximation in L_2 . Formally, this is stated in the following theorem,

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

$$\mathcal{L}[g] = \int_{\mathbb{R} \times \mathbb{X} \times (0,1)} (\beta - \sin \beta) g(\mathbf{x}) \mathcal{N}(d\beta, d\mathbf{x}, du) + \int_{\mathbb{R} \times \mathbb{X} \times (0,1)} \sin \beta g(\mathbf{x}) \tilde{\mathcal{N}}(d\beta, d\mathbf{x}, du) \quad (10)$$

where

$$\mathcal{N} \sim \text{Po}(\nu), \quad \tilde{\mathcal{N}}(d\beta, d\mathbf{x}, du) = \mathcal{N}(d\beta, d\mathbf{x}, du) - \nu(d\beta, d\mathbf{x}, du).$$

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

For any $\epsilon > 0$, construct the approximate Lévy random field \mathcal{L}_ϵ that maps any function $g \in L_2(\mathbb{X}, \pi_\mathbf{x}(d\mathbf{x}))$ to the random variable

$$\mathcal{L}_\epsilon[g] = \int_{\mathbb{R} \times \mathbb{X} \times (0,1)} \beta g(\mathbf{x}) 1_{\{u < (1+\alpha\epsilon^2\beta^{-2})^{-(\alpha+1)/2}\}}(u) \mathcal{N}(d\beta, d\mathbf{x}, du). \quad (11)$$

Then $\mathcal{L}_\epsilon[g] - \mathcal{L}[g]$ converges to 0 in L_2 as ϵ goes to zero, for any $g \in L_2(\mathbb{X}, \pi_\mathbf{x}(d\mathbf{x}))$.

The proof of the theorem is shown in the appendix B, and the squared discrepancy of the continuous approximation is finite:

$$\mathbb{E} \left| \mathcal{L}[g] - \mathcal{L}_\epsilon[g] \right|^2 \leq \|g\|_2^2 \frac{2\gamma}{\pi} \Gamma(\alpha+1) \sin\left(\frac{\pi\alpha}{2}\right) \left(\frac{(1+\alpha)\alpha^{\alpha/2}}{2} + \frac{\alpha^{(2-\alpha)/2}}{2-\alpha} \right) \epsilon^{2-\alpha}. \quad (12)$$

In particular, the squared discrepancy can be calculated exactly when $\alpha = 1$,

$$\mathbb{E} \left| \mathcal{L}[g] - \mathcal{L}_\epsilon[g] \right|^2 = \int_{\mathbb{R} \times \mathbb{X}} g(\boldsymbol{\chi})^2 \frac{\gamma}{\pi} \left(1 - \frac{1}{\beta^2 + \epsilon^2} \right) d\beta d\chi = \gamma \epsilon \|g\|_2^2. \quad (13)$$

This offers guidance on the choice of parameters γ and ϵ , which is further discussed in section 2.4.

2.2 Sparse Representation

In this section, we shall detail the remaining prior distributions for BARK (4) that obtains a sparse representation while selecting the features from the original covariate space.

Denote by \mathbb{X} the support set for kernel location parameter $\boldsymbol{\chi}$. One possible decision is to set $\mathbb{X} = \mathbb{R}^p$, 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.* $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. This reduced the space \mathbb{R}^p to n discrete points. Let $\pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi})$ be a discrete probability measure on \mathbb{X} , which is usually a uniform distribution if no additional information about kernel locations is known before modeling. In order to incorporate the intercept term in the regression seamlessly into this representation, we add an imaginary point \mathbf{x}_0 to \mathbb{X} , such that $\mathbb{X} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, with $K(\mathbf{x}, \mathbf{x}_0) = 1$. It is natural to set the prior distribution for $\boldsymbol{\chi}$ be uniformly over the set of possible kernel locations

$$\{\boldsymbol{\chi}_j\}_{1 \leq j \leq J} \stackrel{iid}{\sim} \text{Un}(\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n\}). \quad (14)$$

This prior specification guarantees that there are at most $n + 1$ distinct kernels in the model (4). When ϵ goes to zero, the prior distribution induces more and more kernels in the expression (4), 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 \mathbf{x}_i , and $J = \sum_{i=0}^n n_i$. The regression mean function can be rewritten as

$$f(\mathbf{x}) = \sum_{i=0}^n \tilde{\beta}_i K(\mathbf{x}, \mathbf{x}_i), \quad \tilde{\beta}_i = \sum_{\{j | \boldsymbol{\chi}_j = \mathbf{x}_i\}} \beta_j.$$

Due to the infinite divisible property of the prior distribution, the prior specification in (9-14) becomes

$$J \sim \text{Po}(\nu_\epsilon^+(\alpha, \gamma, \epsilon)), \quad \mathbf{n} | J \sim \text{MN}(J, \mathbf{1}/(n+1)), \quad \{\tilde{\beta}_i\}_{0 \leq i \leq n} | \mathbf{n} \stackrel{ind}{\sim} t(\alpha, 0, n_i \epsilon^2), \quad (15)$$

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

2.3 Feature Selection

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

$$K(x, \boldsymbol{\chi}) = \exp \left\{ - \sum_{l=1}^p \lambda_l (x_l - \chi_l)^2 \right\}, \quad (16)$$

where the scale parameters λ_l s 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 λ_l is zero, there is no contribution made by the l th variable through the kernel function; on the other hand, if λ_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 (16) 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 λ_l s all equal. Suppose the prior distribution for the sum of all λ_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 $l = 1, \dots, p$, and p is the total number of variables. The exponent term in the Gaussian kernel (16) 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 λ_l s is $\text{Ga}(a_\lambda, b_\lambda)$. Notice that

gamma distribution is infinitely divisible, we can split the total gamma mass equally into individual scale parameters λ_l . Specifically,

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

where $l = 1, \dots, p$, 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, *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 $\boldsymbol{\delta} \in \{0, 1\}^p$ for the scale parameter $\boldsymbol{\lambda}$ 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_λ 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

$$\lambda_l = \begin{cases} \lambda_l^*, & \text{if } \delta_l = 1 \\ 0, & \text{if } \delta_l = 0 \end{cases}, \quad \{\delta_l\}_{1 \leq l \leq p} \overset{iid}{\sim} \text{Bi}(1, p_\lambda), \quad p_\lambda \sim \text{Be}(a_p, b_p). \quad (17)$$

where λ_l^* 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 (17), let

$$\lambda_l^* = \lambda \sim \text{Ga}(a_\lambda, db_\lambda),$$

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

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 (17), let

$$\lambda_l^* \stackrel{iid}{\sim} \text{Ga}(a_\lambda/d, b_\lambda), \quad (18)$$

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_λ and scale b_λ 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 S α S prior specification, $\{\alpha, \gamma, \epsilon\}$. In the continuous approximation, the stable index α serves as the degree of freedom in the student t prior distribution for the regression coefficient β . In particular, $\alpha = 1$ corresponds the Cauchy random field, which induces Cauchy prior distribution on the regression coefficient β . 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.

The approximation threshold ϵ serves as the scale parameter in the student t prior distribution for the regression coefficient β , but both ϵ and the intensity parameter γ determines the number of kernels J in BARK. In the continuous approximation, J has a Poisson prior distribution, with mean γ/ϵ . In addition, it is desirable to control the level of approximation through the L_2 discrepancy (13), such as $\mathbb{E}|\mathcal{L}[g] - \mathcal{L}_\epsilon[g]|^2 \leq 0.05\|g\|_2^2$. In our experience, we chose $\gamma = 1$, $\epsilon = 0.05$, which suggest that on average, we expect to see 20 kernels in the regression, and the square discrepancy of the approximation is no more than 5% of the L_2 norm of the kernel function squared. Generally, one can elicit γ and ϵ via specifying the expected number of kernels through (8), and bounding the L_2 discrepancy in (12).

The spread of the kernel function (16) is controlled by the scale parameters λ_l 's. Since the variables X_l are standardized to have mean 0 and variance 1 before the analysis, if the kernel center $\boldsymbol{\chi}$ 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 \mathbf{X} and $\boldsymbol{\chi}$. When λ_l 's are very close to zero, the kernel function is similar to a point mass at 1; when λ_l 's are very large, the kernel function is similar to a point mass at 0. These cases need to be avoided, because we do not want the kernel behave like the trivial intercept kernel $K(\mathbf{x}, \mathbf{x}_0)$. Four different prior distributions are specified in section 2.3, and we recommend using fixed hyper-parameters $a_\lambda = b_\lambda = 1$, 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 $(x_l - \chi_l)^2/2 \sim \text{Ga}(1/2, 1/2)$, 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 λ_l '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 λ_l 's are different, where d is the number of non-zero λ_l '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-quantile range of the kernel function with $a_\lambda = b_\lambda = 1$ is greater than 0.5 for all interger 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 $(0, 1]$.

In BARK with selection, the probability that each λ_l is non-zero has a Beta prior distribution in (17). This generates a prior distribution for the number of non-zero scale parameters d that corresponds to the Binomial-Beta distribution (see Bernardo and Smith, 1994, pp. 117), with probability mass function

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

We recommend using the uniform hyper-prior distribution, *i.e.* $a_p = b_p = 1$, which induces a discrete uniform prior for the number of non-zero λ_l 's with $\mathbb{P}(d = k) = \frac{1}{p+1}$ for $k = 0, 1, \dots, p$. In other words, the expected number of signal variables in the prior specification is $k/2$. One can also elicit the hyper-parameters from the prior expected number of signal variables, denoted by m . For example, Ley and Steel (2008) suggests fixing $a_p = 1$, and let $b_p = (p - m)/m$.

We put a Gamma prior distribution $\text{Ga}(c, d)$ for the overall precision ϕ . Since ϕ is always in the model, we can set $c = d = 0$, which reduced 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 (4).

2.5 Inference

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

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

where $\mathbf{y} = \{y_1, \dots, y_n\}^T$ and K is the $n \times J$ kernel matrix, with $K_{i,j} = K(\mathbf{x}_i, \boldsymbol{\chi}_j)$.

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

$$p(\phi, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \phi, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}) \pi(\phi, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda} \mid \mathbf{y})}{p(\mathbf{y})}$$

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

$$p(y^* \mid \mathbf{y}) = \int p(y^* \mid \mathbf{x}^*, \phi, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}) p(\phi, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda} \mid \mathbf{y}) d\phi dJ d\boldsymbol{\beta} d\boldsymbol{\chi} d\boldsymbol{\lambda} \quad (19)$$

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 β approximation to the SaS Lévy random field (7), the regression coefficients have independent symmetric Pareto prior distributions with density

$$f(\beta) = \frac{\alpha \epsilon^\alpha}{2} |\beta|^{-\alpha-1} \mathbf{1}_{|\beta| > \epsilon}(\beta).$$

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

Similarly, it is not so convenient to work with t prior distributions (15) on the regression coefficients directly with the alternative approximation. However, we can improve the mixing of the Markov chain by integrating out the regression coefficients and make inference on the precision parameters. To be more specific, decompose the t distribution as a normal mixture of Gamma precisions,

$$\tilde{\beta}_i \stackrel{iid}{\sim} \text{No}(0, n_i \tilde{\varphi}_i^{-1}), \quad \tilde{\varphi}_i \stackrel{iid}{\sim} \text{Ga} \left(\frac{\alpha}{2}, \frac{\alpha \epsilon^2}{2} \right), \quad \text{for } i \in \{i \mid n_i > 0\}.$$

Conditional on the number of kernels J and kernel locations $\{\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_J\}$, we can integrate the regression coefficient $\boldsymbol{\beta}$ and use $\boldsymbol{\varphi}$ to replace the role of $\boldsymbol{\beta}$ in the likelihood function. In the collapsed representation, denote the index of the non-zero elements in \mathbf{n} is $\mathbf{i} = (i_1, \dots, i_m)$, *i.e.* $n_{i_j} > 0$ for $j = 1, \dots, m$. Let $\boldsymbol{\beta}^*$ and $\boldsymbol{\varphi}^*$ be the length- m sub-vector of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\varphi}}$, where

$\beta_j^* = \tilde{\beta}_{i_j}$, $\varphi_j^* = \tilde{\varphi}_{i_j}$. The $n \times (n + 1)$ kernel matrix \tilde{K} is defined by $\tilde{K}_{j,k} = K(\mathbf{x}_j, \mathbf{x}_k)$, where $1 \leq j \leq n$, $0 \leq k \leq n$. Denote by K^* the $n \times m$ sub-matrix of \tilde{K} , where $K_{j,k}^* = K_{j,i_k}$.

After integrating out β^* , the likelihood is

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{n}, \varphi^*, \lambda, \phi) &= \int p(\mathbf{y} \mid \mathbf{n}, \beta^*, \varphi^*, \lambda, \phi) p(\beta^* \mid \varphi^*) d\beta^* \\ &= \frac{\phi^{n/2} \prod_{j=1}^m \left(\frac{\varphi_j^*}{n_{i_j}}\right)^{1/2}}{(2\pi)^{n/2} |\Sigma^*|^{-1/2}} \exp \left\{ -\frac{1}{2} \left(\phi \|\mathbf{y} - K^* \boldsymbol{\mu}^*\|^2 + \sum_{j=1}^m \frac{\varphi_j^*}{n_{i_j}} \mu_j^{*2} \right) \right\}, \end{aligned}$$

where

$$\Sigma^* = (\phi K^{*T} K^* + \text{diag}(\varphi^*))^{-1}, \quad \boldsymbol{\mu}^* = \phi \Sigma^* K^{*T} \mathbf{y}. \quad (20)$$

Denote $\boldsymbol{\theta} = (\mathbf{n}, \varphi^*, \lambda, \phi)$, after integrating out β , instead of sampling from full joint posterior distribution, we only need to sample from $p(\boldsymbol{\theta} \mid \mathbf{y})$. Conditional on $\boldsymbol{\theta}$, the posterior distribution for β^* is $\text{No}(\boldsymbol{\mu}^*, \Sigma^*)$, where $\boldsymbol{\mu}^*$ and Σ^* is defined in (20). Given a new observation \mathbf{x} , suppose $\{\boldsymbol{\theta}^{(m)}\}_{m=1}^M$ are samples of $\boldsymbol{\theta}$ from the posterior MCMC, then $\frac{1}{M} \sum_{m=1}^M f(\mathbf{x} \mid \boldsymbol{\theta}^{(m)}, \boldsymbol{\mu}^{*(m)})$ is a point estimator for the prediction y . This estimator has a smaller variance than $\frac{1}{M} \sum_{m=1}^M f(\mathbf{x} \mid \boldsymbol{\theta}^{(m)}, \beta^{*(m)})$ with $\beta^{*(m)} \sim \text{No}(\boldsymbol{\mu}^{*(m)}, \Sigma^{*(m)})$ due to Rao-Blackwellization.

Because the dimension of $\boldsymbol{\theta}$ is not fixed, we use Reversible Jump Monte Carlo Markov chain (RJ-MCMC) algorithm to sample from the posterior distribution. By integrating out β , we reduced the correlation of $\boldsymbol{\theta}$ and the unknown regression mean function. By sacrificing the conjugacy for φ and β 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,

$$P(y_i = 1 \mid \mathbf{x}_i) = \Phi(f(\mathbf{x}_i)),$$

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

$$y_i = 1(z_i > 0), \quad z_i \stackrel{iid}{\sim} \text{No}(f(\mathbf{x}_i), 1).$$

Conditional on \mathbf{z} , this is exactly the BARK we described in the previous section, except that ϕ is fixed at 1 in the Probit model. Another difference in this specification is that we need to obtain the values of $f(\mathbf{x})$ in order to update \mathbf{z} . Previously, we integrated out β in

the regression for better mixing Markov chains, but now we need to obtain those regression coefficients to calculate the mean function $f(\mathbf{x})$ explicitly. Notice that the conditional posterior distribution for $\boldsymbol{\beta}$ is Normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ as described in (20), the Gibbs sampling is straightforward. After obtaining $f(\mathbf{x})$, we can sample \mathbf{z} from its conditional distribution. If $y_i = 1$, the conditional distribution for z_i is $\text{No}(f(\mathbf{x}_i), 1)$ truncated above zero; if $y_i = 0$, the conditional distribution for z_i is $\text{No}(f(\mathbf{x}_i), 1)$ truncated below zero.

As a result, given a new observation \mathbf{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 $z^{(m)} = f(\mathbf{x} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\beta}^{(k)})$, and $(\boldsymbol{\theta}^{(k)}, \boldsymbol{\beta}^{(k)})$ 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 hyper-parameters are chosen to be $\alpha = 1$, $\epsilon = 0.5$, $\gamma = 10$, $a_\lambda = b_\lambda = a_p = b_p = 1$. 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.

Data Sets	BARK				SVM	BART
	equal	diff	select + equal	select + diff		
Friedman1	7.31	1.22	2.26	1.93	5.36	1.97
Friedman2	1.99	1.07	1.09	1.04	4.36	3.64
Friedman3	3.07	1.46	2.30	1.44	2.70	1.00
Boston Housing	1.44	1.09	1.23	1.20	1.56	1.01
Body Fat	1.39	1.81	1.01	2.19	4.04	1.68
Basketball	1.01	1.01	1.01	1.02	1.16	1.10

Table 1: Predictive mean square errors in regression problems.

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,

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

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

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

The corresponding regression mean functions are

$$\begin{aligned} f_2(\mathbf{x}) &= (x_1^2 + (x_2 x_3 - 1/(x_2 x_4))^2)^{1/2}, \\ f_3(\mathbf{x}) &= \arctan((x_2 x_3 - 1/(x_2 x_4))/x_1). \end{aligned}$$

Independent Gaussian noise with mean 0 and standard deviation 1, 125 and 0.1 are added to the mean function in the three data sets respectively.

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 studied 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 λ 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 λ 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

Boston Housing in BARK with different weights

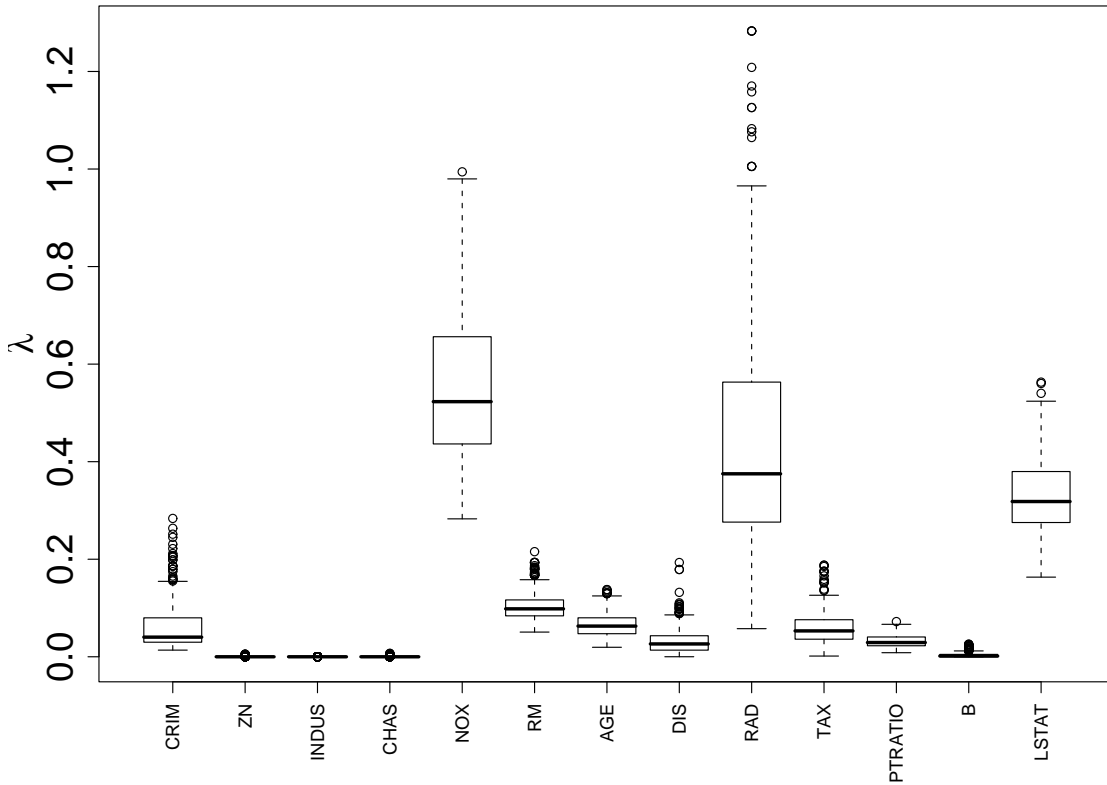


Figure 1: Box plots for the kernel scale parameters in Boston Housing data set in BARK with different weights.

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) contains the data 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 mis-classification 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.

Body Fat in BARK with selection and equal weights

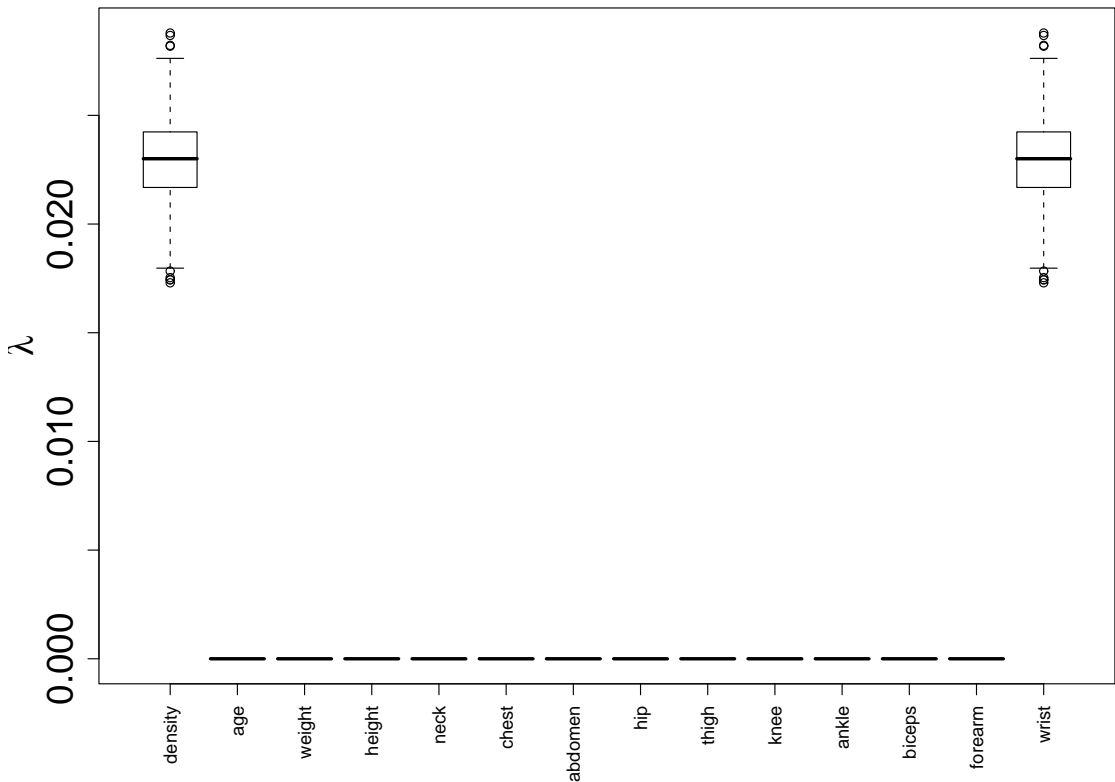


Figure 2: Box plots for the kernel scale parameters in body fat data set in BARK with selection and equal weights.

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, $y = 1_{\{x_1^2 + x_2^2 \leq 2/\pi\}}(\mathbf{x})$. Under this formulation, both class have roughly the same number of sample points.

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

Data Sets	BARK				SVM	BART
	equal	diff	select + equal	select + diff		
Circle 2	1.93%	4.91%	1.88%	1.93%	5.03%	3.97%
Circle 5	13.50%	4.70%	1.47%	1.65%	10.99%	6.51%
Circle 20	49.16%	4.84%	2.09%	3.69%	44.10%	15.10%
Bank	1.05%	1.25%	0.55%	0.88%	1.12%	0.50%
WDBC	2.70%	4.02%	2.49%	6.09%	2.70%	3.36%
Ionosphere	5.33%	8.59%	5.78%	10.87%	5.17%	7.34 %

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

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) 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 suggest that BARK with selection and equal weights has a very good out-of-sample prediction. However, the posterior probability that $\lambda_l > 0$ 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, and Ouyang *et al.* (2008) extends BARK with lower rank models to capture those structures.

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

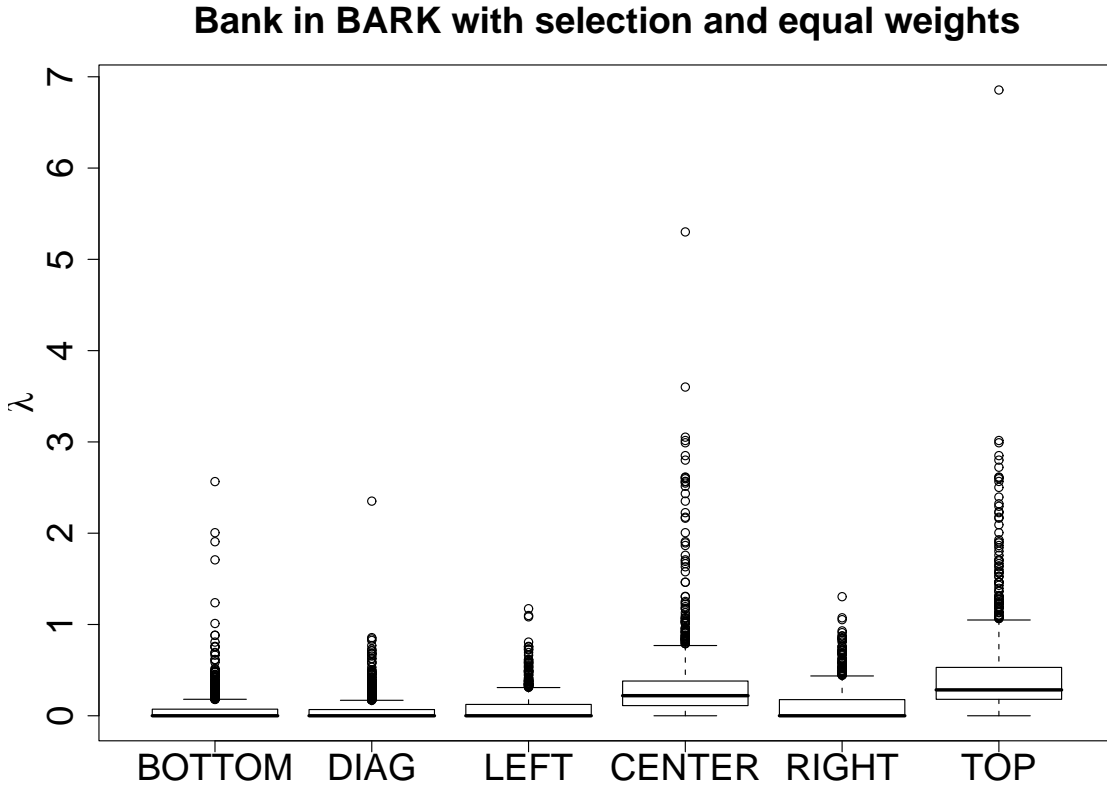


Figure 3: Box plots for the kernel scale parameters in Swiss bank note data set in BARK with selection and equal weights.

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 non-parametric 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 to 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 ν to $\mathbb{R} \times \mathbb{X} \times \Lambda$, where Λ 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 Λ 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 $-\infty = c_0 < c_1 = 0 < c_2 < \dots < c_{d-1} < c_d = \infty$. 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.

Appendices

A Poisson Construction for Lévy Random Field

Generally, when a Lévy measure ν satisfies the L_2 local integrability, *i.e.*

$$\iint_{\mathbb{R} \times \mathbb{X}} (1 \wedge \beta^2) \nu(d\beta, d\boldsymbol{\chi}) < \infty \quad (21)$$

and $\nu(\{0\}, \mathbb{X}) = 0$, it induces a random measure \mathcal{L} which assigns independent infinitely-divisible random variables $\mathcal{L}(A_i)$ to disjoint Borel sets $A_i \in \mathbb{X}$, with characteristic functions

$$\mathbb{E} [e^{it\mathcal{L}(A)}] = \exp \left\{ it\delta_h(A) + \iint_{\mathbb{R} \times A} (e^{it\beta} - 1 - it h(\beta)) \nu(d\beta, d\boldsymbol{\chi}) \right\}. \quad (22)$$

h is called a compensator function, which is bounded and $O(\beta^2)$ for $\beta \approx 0$. It makes the integrand in (22) bounded, and guarantees the characteristic functions are well defined. The signed measure $\delta_h(d\boldsymbol{\chi})$ on \mathbb{X} in (22) is uniquely determined by the compensator function h . For example, in the $S\alpha S$ case, if we set $h(\beta) = \sin(\beta)$, then corresponding drift term δ_h is 0, and the characteristic function (22) is simplified to $\exp \{-\gamma\pi(A) |t|^\alpha\}$.

When the Lévy measure ν satisfies the stronger L_1 local integrability condition, *i.e.*

$$\iint_{\mathbb{R} \times \mathbb{X}} (1 \wedge |\beta|) \nu(d\beta, d\boldsymbol{\chi}) < \infty,$$

no compensation is required, and the characteristic functions (22) is well defined with $h \equiv 0$. For example, when the stable index $0 < \alpha < 1$ in the $S\alpha S$ Lévy measure, L_1 local integrability condition is satisfied.

Suppose the Lévy measure $\nu(d\beta d\boldsymbol{\chi}) = \pi_\beta(\beta) \pi(\boldsymbol{\chi})$ is separable, where $\pi(\boldsymbol{\chi})$ is a probability measure on \mathbb{X} . Denote by $L_2(\mathbb{X}, \pi(d\boldsymbol{\chi}))$ the linear space of all Borel measurable functions that is square integrable with respect to $\pi(d\boldsymbol{\chi})$, *i.e.*

$$L_2(\mathbb{X}, \pi(d\boldsymbol{\chi})) = \left\{ \int_{\mathbb{X}} g^2(\boldsymbol{\chi}) \pi(d\boldsymbol{\chi}) < \infty. \right\}.$$

Notice that the Gaussian kernel $K(\mathbf{x}, \boldsymbol{\chi})$ defined in (16) is in $L_2(\mathbb{X}, \pi(d\boldsymbol{\chi}))$ for any given \mathbf{x} .

A Lévy random measure induces a random field $\mathcal{L} : g \mapsto \mathcal{L}[g]$ which maps functions $g \in L_2(\mathbb{X}, \pi(d\boldsymbol{\chi}))$ to random variables $\mathcal{L}[g] = \int_{\mathbb{X}} g(\boldsymbol{\chi}) \mathcal{L}(d\boldsymbol{\chi})$. The characteristic function for $\mathcal{L}[g]$ is

$$\mathbb{E} [e^{it\mathcal{L}[g]}] = \exp \left\{ \iint_{\mathbb{R} \times \mathbb{X}} (e^{it\beta g(\boldsymbol{\chi})} - 1 - it h(\beta) g(\boldsymbol{\chi})) \nu(d\beta, d\boldsymbol{\chi}) \right\}. \quad (23)$$

It can be constructed from a compensated Poisson random field. Begin with the Poisson random measure $\mathcal{N} \sim \text{Po}(\nu)$ on $\mathbb{R} \times \mathbb{X}$, denote the centered Poisson random measure

$\tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}) = \mathcal{N}(d\beta, d\boldsymbol{\chi}) - \nu(d\beta, d\boldsymbol{\chi})$, which induces an isometry from $L_2(\mathbb{R} \times \mathbb{X}, \nu(d\beta, d\boldsymbol{\chi}))$ to the square-integrable zero-mean random variables, (see [Sato, 1999](#), page 38). Now set

$$\mathcal{L}[g] = \iint_{\mathbb{R} \times \mathbb{X}} (\beta - h(\beta))g(\boldsymbol{\chi})\mathcal{N}(d\beta, d\boldsymbol{\chi}) + \iint_{\mathbb{R} \times \mathbb{X}} h(\beta)g(\boldsymbol{\chi})\tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi})$$

for any $g \in L_2(\mathbb{X}, \pi(d\boldsymbol{\chi}))$. Then $\mathcal{L}[g]$ is a SoS random variable with characteristic function [\(23\)](#).

B Proof of Theorem 1

For any $g \in L_2(\mathbb{X}, \pi(d\boldsymbol{\chi}))$, rewrite the Poisson construction [\(10\)](#) as

$$\begin{aligned} \mathcal{L}[g] &= \int_{(-1,1) \times \mathbb{X} \times (0,1)} \beta g(\boldsymbol{\chi})\tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}, du) + \int_{(-1,1) \times \mathbb{X} \times (0,1)} (\beta - \sin \beta)g(\boldsymbol{\chi})\nu(d\beta, d\boldsymbol{\chi}, du) + \\ &\int_{(-1,1)^c \times \mathbb{X} \times (0,1)} \beta g(\boldsymbol{\chi})\mathcal{N}(d\beta, d\boldsymbol{\chi}, du) - \int_{(-1,1)^c \times \mathbb{X} \times (0,1)} \sin \beta g(\boldsymbol{\chi})\nu(d\beta, d\boldsymbol{\chi}, du). \end{aligned} \quad (24)$$

Because $\beta - \sin \beta = O(\beta^2)$ when $\beta \approx 0$, and $1_{|\beta| < 1}(\beta)\beta^2 g(\boldsymbol{\chi})$ is ν -integrable, the second term in [\(24\)](#) is finite. In addition, $\beta - \sin \beta$ is an odd function, and $\nu(d\beta, d\boldsymbol{\chi})$ is symmetric about zero on the first dimension, so $\int_{(-1,1) \times \mathbb{X} \times (0,1)} (\beta - \sin \beta)g(\boldsymbol{\chi})\nu(d\beta, d\boldsymbol{\chi})du = 0$. Similarly, the fourth terms in [\(24\)](#) is also zero.

Notice that the difference between [\(24\)](#) and [\(11\)](#) is

$$\begin{aligned} \mathcal{L}[g] - \mathcal{L}_\epsilon[g] &= \int_{(-1,1) \times \mathbb{X} \times (0,1)} 1_{\{(1+\alpha\epsilon^2\beta^{-2})^{-(\alpha+1)/2} \leq u < 1\}}(u)\beta g(\boldsymbol{\chi})\tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}, du) + \\ &\int_{(-1,1)^c \times \mathbb{X} \times (0,1)} 1_{\{(1+\alpha\epsilon^2\beta^{-2})^{-(\alpha+1)/2} \leq u < 1\}}(u)\beta g(\boldsymbol{\chi})\mathcal{N}(d\beta, d\boldsymbol{\chi}, du). \end{aligned}$$

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

$$\begin{aligned} \mathbb{E} \left| \mathcal{L}[g] - \mathcal{L}_\epsilon[g] \right|^2 &= \int_{\mathbb{R} \times \mathbb{X} \times (0,1)} 1_{\{(1+\alpha\epsilon^2\beta^{-2})^{-(\alpha+1)/2} \leq u < 1\}}(u)\beta^2 g^2(\boldsymbol{\chi})\nu(d\beta, d\boldsymbol{\chi}, du) \\ &= \|g\|_2^2 \frac{2\gamma\Gamma(\alpha+1)}{\pi} \sin\left(\frac{\pi\alpha}{2}\right) \int_0^\infty \left(1 - \left(1 + \frac{\alpha\epsilon^2}{\beta^2}\right)^{-\frac{1+\alpha}{2}}\right) \beta^{1-\alpha} d\beta \end{aligned} \quad (25)$$

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

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

Therefore,

$$\int_{\delta}^{\infty} \left(1 - \left(1 + \frac{\alpha \epsilon^2}{\beta^2} \right)^{-(1+\alpha)/2} \right) \beta^{1-\alpha} d\beta \leq \int_{\delta}^{\infty} \frac{\alpha(1+\alpha)\epsilon^2}{2\beta^{1+\alpha}} d\beta = \frac{(1+\alpha)\epsilon^2}{2\delta^\alpha} = \frac{(1+\alpha)\alpha^{\alpha/2}\epsilon^{2-\alpha}}{2}.$$

In addition,

$$\int_0^{\delta} \left(1 - \left(1 + \frac{\alpha \epsilon^2}{\beta^2} \right)^{-(1+\alpha)/2} \right) \beta^{1-\alpha} d\beta \leq \int_0^{\delta} \beta^{1-\alpha} d\beta = \frac{\delta^{2-\alpha}}{2-\alpha} = \frac{\alpha^{(2-\alpha)/2}\epsilon^{2-\alpha}}{2-\alpha}.$$

Combining these two bounds into (25),

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

Because $0 < \alpha < 2$, the above term goes to zero as ϵ approaches zero. In conclusion, $\mathcal{L}[g] - \mathcal{L}_\epsilon[g]$ converges to 0 in L_2 for any $g \in L_2(\mathbb{X}, \pi(d\boldsymbol{\chi}))$.

C 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 $\{J, \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \phi\}$ reduced to $\{\mathbf{n}, \boldsymbol{\beta}^*, \boldsymbol{\varphi}^*, \boldsymbol{\lambda}, \phi\}$ in the collapsed representation. Since $\boldsymbol{\beta}^*$ is integrated out in the likelihood, we only need to sample $(\mathbf{n}, \boldsymbol{\varphi}^*, \boldsymbol{\lambda}, \phi \mid \mathbf{y})$.

1. Update $(\mathbf{n}, \boldsymbol{\varphi}^*)$ using RJ-MCMC algorithm, conditional on other parameters.
2. Update $\boldsymbol{\lambda}$ using standard Metropolis-Hasting algorithm, conditional on other parameters. Pick one of λ_l , say $\lambda_{l^{(prop)}}$ from vector $\boldsymbol{\lambda}$ at random, and update it via normal random walk on the log scale to $\lambda_{l^{(prop)}}^{(prop)}$. The acceptance rate is the minimal of 1 and

$$\frac{p(\mathbf{y} \mid \boldsymbol{\theta}^{(prop)}) \pi_{\lambda}(\lambda_{l^{(prop)}}^{(prop)}) \lambda_{l^{(prop)}}^{-1}}{p(\mathbf{y} \mid \boldsymbol{\theta}) \pi_{\lambda}(\lambda_{l^{(prop)}}) \lambda_{l^{(prop)}}^{(prop)-1}}.$$

3. Update ϕ using standard Metropolis-Hasting algorithm, conditional on other parameters. The prior density for ϕ is proportional to ϕ^{-1} , which cancels the proposal density, hence the acceptance rate is the minimal of 1 and

$$\frac{p(\mathbf{y} \mid \boldsymbol{\theta}^*) \phi^{*-1} \phi^{-1}}{p(\mathbf{y} \mid \boldsymbol{\theta}) \phi^{-1} \phi^{*-1}} = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}^*)}{p(\mathbf{y} \mid \boldsymbol{\theta})}.$$

For classification problems,

1. Update $(\mathbf{n}, \boldsymbol{\varphi}^*, \boldsymbol{\lambda})$ as in the regression case (1-2), conditional on the latent normal random variable \mathbf{z} . Notice that $\phi \equiv 1$ in the classification case.
2. Simulate $\boldsymbol{\beta}^* \sim \text{No}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ is defined in (20).
3. Simulate \mathbf{z} from its full conditional distribution given $\mathbf{y}, \mathbf{n}, \boldsymbol{\beta}^*$, *i.e.*

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

Now we detail the RJ-MCMC algorithm. Suppose the current $\boldsymbol{\theta}$ 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 $\boldsymbol{\chi}_j^{(prop)}$ uniformly from $\{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n\}$, say $\boldsymbol{\chi}_j^{(prop)} = \mathbf{x}_{i^{(prop)}}$. If there is already some kernel located at $\boldsymbol{\chi}_j^{(prop)}$, or $n_{i^{(prop)}} \neq 0$, update $n_{i^{(prop)}}^{(prop)} = n_{i^{(prop)}} + 1$, and keep $\boldsymbol{\varphi}^*$ unchanged. If no existing kernel located at $\boldsymbol{\chi}_j^{(prop)}$, or $n_{i^{(prop)}} = 0$, update $n_{i^{(prop)}}^{(prop)} = 1$, propose a new regression coefficient precision from the prior distribution $\varphi^* \sim (\alpha/2, \alpha\epsilon^2/2)$, and add it into the current $\boldsymbol{\varphi}^*$.

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 δ of its regression precision. In other words, the probability of selecting kernel at location \mathbf{x}_i to kill is proportional to $n_i \tilde{\varphi}_i^\delta$. 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 minimal of 1 and the product of the conditional posterior density ratio and the proposal density ratio,

$$\frac{p(y | \boldsymbol{\theta}^{(prop)})\pi(\boldsymbol{\theta}^{(prop)})q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(prop)})}{p(y | \boldsymbol{\theta})\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}^{(prop)} | \boldsymbol{\theta})} = \frac{p(y | \boldsymbol{\theta}^{(prop)})\nu^+(\alpha, \gamma, \epsilon)p_d(k+1)p^{(kill)}}{p(y | \boldsymbol{\theta})n_{i^{(prop)}}^{(prop)}p_b(k)}.$$

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

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

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

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

$$\frac{p(y | \boldsymbol{\theta}^{(prop)}) \text{Ga}(\varphi_{i^{(prop)}}^{*(prop)} | \alpha/2, \alpha\epsilon^2/2) \varphi_{i^{(prop)}}^{*(prop)}}{p(y | \boldsymbol{\theta}) \text{Ga}(\varphi_{i^{(prop)}}^* | \alpha/2, \alpha\epsilon^2/2) \varphi_{i^{(prop)}}^*} \quad (26)$$

- (b) **Update (\mathbf{n}, φ) .** 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 φ^* , otherwise, propose a new φ^* from its prior distribution. The multi-nominal prior density for \mathbf{n} exactly cancels the proposal density, hence the acceptance rate is the minimal of 1 and

$$\frac{p(y | \boldsymbol{\theta}^{(prop)})}{p(y | \boldsymbol{\theta})}. \quad (27)$$

References

- Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*, **88**, 669–679.
- Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. John Wiley & Sons.
- Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, (ed. D. Haussler), pp. 144–152.
- Breiman, L. (1996) Bagging predictors. *Machine Learning*, pp. 123–140.
- Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.*, **80**, 580–598.
- Chakraborty, S., Ghosh, M. and Mallick, B. K. (2004) Bayesian nonlinear regression for large p small n problem. Tech. Rep. 2004-01, University of Florida Department of Statistics.
- Chipman, H. A., George, E. I. and McCulloch, R. E. (2007) Bayesian ensemble learning. In *Advances in Neural Information Processing Systems 19*, (eds. B. Schölkopf, J. C. Platt and T. Hoffman), pp. 265–272. Cambridge, MA: MIT Press.
- Clyde, M. and George, E. I. (2004) Model uncertainty. *Statistical Science*, **19**, 81–94.
- Clyde, M. A., House, L. L., Tu, C. and Wolpert, R. L. (2005) Bayesian nonparametric function estimation using overcomplete kernel representations and Lévy random field priors. In *Statistische und Probabilistische Methoden der Modellwahl*, vol. 2, (eds. J. O. Berger, H. Dette, G. Lugosi and A. Munk), pp. 2628–2632.
- Clyde, M. A., House, L. L. and Wolpert, R. L. (2006) Nonparametric models for proteomic peak identification and quantification. In *Bayesian Inference for Gene Expression and Proteomics*, (eds. K.-A. Do, P. Müller and M. Vannucci), pp. 293–308. Cambridge, UK: Cambridge Univ. Press.
- Clyde, M. A. and Wolpert, R. L. (2007) Nonparametric function estimation using overcomplete dictionaries. In *Bayesian Statistics 8*, (eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 91–114. Oxford, UK: Oxford Univ. Press.
- Cont, R. and Tankov, P. (2004) *Financial modelling with jump processes*. London, UK: Chapman & Hall/CRC.
- Cook, R. D. and Lee, H. (1999) Dimension reduction in binary response regression. *J. Am. Stat. Assoc.*, **94**, 1187–1200.

- Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge, UK: Cambridge Univ. Press.
- Flury, B. and Riedwyl, H. (1988) *Multivariate Statistics: a Practical Approach*. Chapman and Hall Ltd.
- Friedman, J. H. (1991) Multivariate adaptive regression splines (Disc: P67-141). *Ann. Stat.*, **19**, 1–67.
- George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–374.
- Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Harrison, D. and Rubinfeld, D. L. (1978) Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics Management*, **5**, 81–102.
- Hofmann, T., Schölkopf, B. and Smola, A. J. (2008) Kernel methods in machine learning. *The Annals of Statistics*, **36**, 1171–1220.
- Johnson, R. (1995) CMU StaLib Datasets Archive. On-line at <http://lib.stat.cmu.edu/datasets/>.
- Ley, E. and Steel, M. F. (2008) On the effect of prior assumptions in bayesian model averaging with applications to growth regression. CRISM Working Paper 07-08, University of Warwick.
- Newman, D. J., Hettich, S., Blake, C. L. and Merz, C. J. (1998) UCI repository of machine learning databases. On-line at <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
- Nott, D. J. and Kohn, R. (2005) Adaptive sampling for Bayesian variable selection. *Biometrika*, **92**, 747–763.
- Ouyang, Z., Clyde, M. A. and Wolpert, R. L. (2008) Fully bayesian low rank kernel models for classification. Discussion paper, Duke University, Department of Statistical Science.
- Pillai, N. S., Wu, Q., Liang, F., Mukherjee, S. and Wolpert, R. L. (2007) Characterizing the function space for bayesian kernel models. *Journal of Machine Learning Research*, **8**, 1769–1797.
- Samorodnitsky, G. and Taqqu, M. S. (1994) *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, vol. 1 of *Stochastic Modeling Series*. New York, NY: Chapman & Hall.
- Sato, K.-i. (1999) *Lévy Processes and Infinitely Divisible Distributions*. Cambridge, UK: Cambridge Univ. Press.

- Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. Springer-Verlag.
- Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–343.
- Tipping, M. E. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, **1**, 211–244.
- Tu, C., Clyde, M. A. and Wolpert, R. L. (2006) Lévy adaptive regression kernels. Discussion Paper 2006-08, Duke University ISDS.
- Wolberg, W. H., Street, W. N. and Mangasarian, O. L. (1995) UCI repository of machine learning databases. On-line at <http://www.ics.uci.edu/~mllearn/MLRepository.html>.