BAYESIAN ADDITIVE REGRESSION KERNELS

by

Zhi Ouyang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Merlise A. Clyde, Supervisor

_____
Dr. Robert L. Wolpert, Supervisor

_____
Dr. David B. Dunson

_____
Dr. Mark Huber

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2008

ABSTRACT

BAYESIAN ADDITIVE REGRESSION KERNELS

by

Zhi Ouyang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Dr. Merlise A. Clyde, Supervisor

_____
Dr. Robert L. Wolpert, Supervisor

_____
Dr. David B. Dunson

_____
Dr. Mark Huber

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2008

Abstract

We propose a general Bayesian "sum of kernels" model, named Bayesian Additive Regression Kernels (BARK), for both regression and classification problems. The unknown mean function is represented as a weighted sum of kernel functions, which is constructed by a prior using symmetric α-stable ($S\alpha S$) Lévy random fields. Both truncation and continuous approximations for the $S\alpha S$ Lévy random fields are investigated, which lead to specifications of joint prior distributions for the number of kernel functions, the regression coefficients and kernel parameters.

Dimension reduction and variable selection techniques are commonly used for gaining insight and making more accurate predictions in the regression and classification problems. We detail a fully Bayesian approach for dimension reduction through low rank Gaussian kernel functions in the non-parametric kernel logistic regression model. We also demonstrate a direct feature selection procedure which facilitates a hierarchical mixture prior distribution of point mass at zero and a gamma distribution on the kernel scale parameters. Finally, we present some preliminary investigation adding dependence structure into this feature selection procedure through a Markov model on the kernel scale parameters. Reversible jump algorithms are implemented to facilitate the posterior inference, and the methods are illustrated with several simulated data sets and real data sets.

Acknowledgements

I would like to take this opportunity to acknowledge with gratitude my friends and colleagues who have helped and supported me over the past four years. First and foremost, I would like to express my deep and sincere gratitude to my fabulous advisors, Dr. Merlise Clyde and Dr. Robert Wolpert, for their patience, encouragement and guidance throughout the development of this dissertation. I have been fortunate to work with them, and their broad knowledge and enthusiasm for sciences have always been my inspirations.

I also wish to thank my committee members, Dr. David Dunson and Dr. Mark Huber for their insightful comments. I am grateful to Dr. Mike West, Dr. James Berger, Dr. Alan Gelfand, Dr. Ian Dinwoodie, Dr. Woncheol Jang, Dr. Yuguo Chen and Dr. Feng Liang, for teaching me various aspects of statistics during my Ph.D. study, as well as Dr. David Banks, Dr. Dalene Stangl and Dr. Jerry Reiter, for their support and help throughout the years. I am also very thankful to Dr. Diane Lambert and Dr. Daryl Pregibon for their mentoring and personal help during my internship at Google.

I am extremely grateful for the friends I made at Duke. In particular, I would like to thank Huiyan, Liang, Robin and Joyee, for accompanying me through the first year exam, the prelim, and the defense; Fei, Yuhong, Quanlin, for sharing apartments with me; Jarad, Eric, Nateshi and many others for discussions during my thesis writing. I also thank the wonderful DSS staff, Kris, Karen, Pat, Nikki, Anne, Susan, Eric and Lance for their indispensable help.

I would like to dedicate this dissertation to my family, Mom, Dad and my wife Ying Yang, for their endless support and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistical models are used to understand the relationship between a response variable and the explanatory variables, or covariates. When the response variable is continuous, it is often referred as regression problems; when the response variable is discrete, it is referred as classification problems. Bayesian Additive Regression Kernels (BARK) provides a unified approach for both the regression and the classification problems.

There are two major goals in those problems: making predictions for a set of future observations, and learning the underlining structure of how the explanatory variables influence the response. The prediction task can be done with a flexible model for the mean function. Typically, people used "sum of generating functions" to capture both the additive and the interactive effects from the explanatory variables. In this thesis, we shall focus on "sum of kernels" models, and compare its predictive performance with the classical kernel model Support Vector Machines (SVM), and the "sum of trees" model Bayesian Additive Regression Trees (BART). Making inference for the kernel functions reveals the relationship of the explanatory variables and the response. For example, with low rank Gaussian kernel functions, the model effectively projects the original covariate space to a dimension reduced subspace; while shrinking the diagonal elements to zero in the diagonal Gaussian kernels selects out those variables in the regression model.

The mean function in BARK is a stochastic integration of the kernel function, whose prior distributions are specified through random measures. In section 1.1, we define the terminology and review Lévy random measures. In section 1.2, we

present a brief review of Lévy random fields, and its construction using Poisson random measures. In section 1.3, we present a brief review of symmetric α-stable Lévy random measures and the corresponding Lévy random fields. In section 1.4, we outline the structure of the remaining chapters in this thesis.

1.1 Lévy Random Measures

A Lévy random measure $\mathcal{L}$ on Ω assigns independent infinitely-divisible random variables $\mathcal{L}(A_i)$ to disjoint Borel sets $A_i \in \Omega$. In general, the characteristic function for any infinitely-divisible random variable has the following form

$$\begin{aligned} \mathsf{E}\left[e^{it\mathcal{L}(A)}\right] &= \exp\Big\{it\delta_h(A) - \frac{1}{2}t^2\Sigma(A) + \\ &\qquad \iint_{\mathbb{R}\times A}\left(e^{it\beta} - 1 - ith(\beta)\right)\nu(d\beta, d\omega)\Big\}, \end{aligned} \tag{1.1}$$

where δ_h is a finite signed measure on Ω, Σ is a positive measure on Ω (Khinchine and Lévy, 1936; Rajput and Rosiński, 1989). The σ-finite measure ν on $\mathbb{R}\times\Omega$ is called the Lévy measure, which satisfies the L_2 local integrability condition

$$\iint_{\mathbb{R}\times K}(1\wedge\beta^2)\nu(d\beta, d\omega) < \infty$$

for each compact $K \in \Omega$ and $\nu(\{0\}, \Omega) = 0$, for more details see Jacod and Shiryaev (1987, *p.* 75), Cont and Tankov (2004, *pp.* 457-459).

The compensator function h in (1.1) is determined uniquely by the characteristic triplet (δ_h, Σ, ν). When the Lévy measure satisfies the stronger L_1 local integrability condition

$$\iint_{\mathbb{R}\times K}(1\wedge\beta)\nu(d\beta, d\omega) < \infty$$

for each compact $K \in \Omega$, we can set $h \equiv 0$ with appropriate δ_h. This non-negative Lévy random measure is called a "completely random measure" by Kingman (1967). If the Lévy measure is only L_2 local integrable, but not L_1 local integrable, in order to make the integrand in (1.1) bounded and $O(\beta^2)$ near $\beta \approx 0$, the compensator function must be a bounded measurable function satisfying

$$h(\beta) = \beta + O(\beta^2), \quad \beta \approx 0.$$

Since the random measure $\mathcal{L}$ maybe written as the sum of a Gaussian part, which assigns independent distributed normal random variables with mean $\delta_h(A_i)$ and variance $\sigma(A_i)$ to disjoint sets A_i, and a remaining part, which assigns independent random variables with characteristic function

$$\exp\left\{\iint_{\mathbb{R}\times A}\left(e^{it\beta} - 1 - ith(\beta)\right)\nu(d\beta, d\omega)\right\}. \tag{1.2}$$

We call a random signed measure $\mathcal{L}$ without the Gaussian component a Lévy random measure. In chapter 2, we used the Lévy random measure on $\Omega = \mathbb{X}\times\mathbb{R}_+^r$, the product space for the kernel location and scale parameters, and in chapter 3, we used the Lévy random measure on $\Omega = \mathbb{X}$, the kernel location space only.

1.2 Lévy Random Fields

The Lévy random measure that maps Borel set A to the random variable with characteristic function (1.2) can be viewed as a linear operator that maps the indicator function $1_A(\omega)$ to the same random variable. In general, the Lévy random measure induces a linear mapping from functions $g : \Omega \to \mathbb{R}$ to random variables $\mathcal{L}[g]$, with

following characteristic function

$$\mathsf{E}\left[e^{it\mathcal{L}[g]}\right] = \exp\left\{\iint_{\mathbb{R}\times\Omega}\left(e^{it\beta g(\omega)} - 1 - ith(\beta)g(\omega)\right)\nu(d\beta, d\omega)\right\}. \tag{1.3}$$

Such a mapping is called a Lévy random field. It is straight forward to verify that the domain of the Lévy random field includes all simple functions, thus by continuity, it at least includes all bounded measurable compactly-supported functions, see (Wolpert and Taqqu, 2005), or (Rajput and Rosiński, 1989; Kwapień and Woyczyński, 1992, *p.* 9) for a general discussion on all possible functions g that the integral in 1.3 are well defined. This includes kernel functions, in particular, the Gaussian kernel functions, we are going to use in the later chapters.

The key for making tractable Bayesian posterior inference using Lévy random fields is through the construction based on Poisson random fields. Specifically, when ν satisfies the L_1 local integrability condition, we may set the compensator $h \equiv 0$. Denote by the Poisson random measure $\mathcal{N}(d\beta, d\omega) \sim \mathsf{Po}(\nu(d\beta, d\omega))$ on $\mathbb{R}\times\Omega$, which assigns independent Poisson random variables with $\mathsf{Po}(\nu(B_i))$ distributions to disjoint Borel sets $B_i \subset \mathbb{R}\times\Omega$. For any bounded measurable compactly-supported function $g(\omega)$, the random variable induced by the Lévy random field can be constructed as follows

$$\mathcal{L}[g] = \iint_{\mathbb{R}\times\Omega}\beta g(\omega)\mathcal{N}(d\beta, d\omega) = \sum_{j=0}^{J}\beta_j g(\omega_j),$$

where $\{(\beta_j, \omega_j)\}$ is the set of $J \leq \infty$ support points in $\mathcal{N}(d\beta, d\omega)$.

When $\nu(d\beta, d\omega)$ is not L_1 locally integrable, but L_2 locally integrable, we need to take care of the compensator function in the Poisson construction. Let $\mathcal{N} \sim \mathsf{Po}(\nu)$ be the same Poisson random measure introduced above, denote by the centered Poisson

random measure $\tilde{\mathcal{N}}(d\beta, d\omega) = \mathcal{N}(\nu(d\beta, d\omega) - \nu(d\beta, d\omega)$, which induces an isometry from $L_2(\mathbb{R} \times \Omega, \nu(d\beta, d\omega))$ to square-integrable zero-mean random variables (Sato, 1999, *pp.* 38). The random variable induced by the Lévy random field can be constructed from

$$\mathcal{L}[g] = \iint_{\mathbb{R}\times\Omega} [\beta - h(\beta)]\, g(\omega)\mathcal{N}(d\beta, d\omega) + \iint_{\mathbb{R}\times\Omega} h(\beta)\, g(\omega)\tilde{\mathcal{N}}(d\beta, d\omega) \tag{1.4}$$

for any measurable function g that (1.4) converges.

1.3 Symmetric α-Stable Lévy Random Fields

In general, a random variable X is said to have a stable distribution $\mathsf{St}(\alpha, \beta_s, \gamma, \delta)$ if its characteristic function has the following form

$$\mathsf{E}\left[e^{itX}\right] = \begin{cases} \exp\left\{i\delta t - \gamma|t|^\alpha - i\beta_s\gamma \tan\frac{\pi\alpha}{2}\left(t - |t|^\alpha \mathrm{sgn}t\right)\right\} & \text{if } \alpha \neq 1 \\ \exp\left\{i\delta t - \gamma|t| - \frac{2i\beta_s\gamma t}{\pi}\log|t|\right\} & \text{if } \alpha = 1 \end{cases}$$

where $1 < \alpha \leq 2$ is the stable index, $1 \leq \beta_s \leq 1$ is the skewness parameter, $\gamma > 0$ is the intensity parameter, and $\delta \in \mathbb{R}$ is the drift parameter, see Samorodnitsky and Taqqu (1994, *pp.* 113-117) for more details. In this thesis, we shall focus on the case that $0 < \alpha < 2$ and $\beta_s = \delta = 0$, which is called the symmetric α-stable ($S\alpha S$) distribution, with characteristic function

$$\mathsf{E}\left[e^{itX}\right] = \exp\left\{-\gamma|t|^\alpha\right\} = \int_{\mathbb{R}} \left(e^{itu} - 1 - it\sin u\right)\nu(du), \tag{1.5}$$

where $\nu(du) = \frac{\alpha\gamma}{\pi}\Gamma(\alpha)\sin\frac{\pi\alpha}{2}|u|^{-1-\alpha}\,du$ is called the $S\alpha S$ Lévy measure. In particular, when $\alpha = 1$, X has Cauchy distribution, with probability density function $(\pi\gamma)^{-1}(1+x^2/\gamma^2)^{-1}$. The function $\sin u$ serves as the role of a compensator, making the integrand

in (1.5) finite. In addition, the random measure constructed from this compensator function with drift $\delta = 0$ does not involve a Gaussian component, which is indeed a Lévy random measure described in section 1.1.

Specifically, to construct a $S\alpha S$ Lévy random measure on Ω, we start from Lévy measure $\nu(d\beta, d\omega)$ on $\mathbb{R} \times \Omega$

$$\nu(d\beta, d\omega) \quad = \quad \frac{\alpha\gamma}{\pi}\Gamma(\alpha)\sin\frac{\pi\alpha}{2}|\beta|^{-1-\alpha}d\beta\pi_\omega(d\omega),$$

where $\pi_\omega(d\omega)$ is a σ-finite measure on Ω. In this thesis, we will use probability measure $\pi_\omega(d\omega)$ in the regression and classification problems. This induces a $S\alpha S$ Lévy random measure mapping disjoint Borel sets $A_i \in \Omega$ to independent $S\alpha S$ random variables $\mathcal{L}(A_i) \sim \mathsf{St}(\alpha, 0, \gamma\pi_\omega(A_i), 0)$.

Furthermore, the corresponding $S\alpha S$ Lévy random field maps function $g : \Omega \to \mathbb{R}$ to $S\alpha S$ random variable $\mathcal{L}[g] \sim \mathsf{St}(\alpha, 0, \gamma^*, 0)$, where $\gamma^* = \gamma\int_\Omega |g(\omega)|^\alpha\pi_\omega(d\omega)$. With the compensator function $h(\beta) = \sin\beta$, the random variable can be constructed from (1.4), or

$$\mathcal{L}[g] = \iint_{\mathbb{R}\times\Omega}[\beta - \sin\beta]\, g(\omega)\mathcal{N}(d\beta, d\omega) + \iint_{\mathbb{R}\times\Omega}\sin\beta\, g(\omega)\tilde{\mathcal{N}}(d\beta, d\omega).$$

1.4 Overview

The remainder of this thesis investigates various models that use $S\alpha S$ Lévy random fields as the prior distribution for the unknown mean function in both regression and classification problems.

Chapter 2 focus on problems where a lower dimensional structure is embedded in the original covariates. Unlike traditional dimension reduction techniques, we

propose a fully Bayesian approach to make joint inference for both making dimension reduction and learning the unknown regression mean function. The low dimension structure is captured by a common eigenvector matrix in the low rank Gaussian kernel functions, which induces a projection mapping from the original covariate space to the dimension reduction space. We focused on binary classification models, and built a non-parametric logistic kernel regression model. We introduced a $S\alpha S$ Lévy random measure for both the kernel location parameters and the kernel eigenvalues, which induces a mapping from the kernel function to a $S\alpha S$ random variable for modeling the logistic regression function. Truncation approximation for the $S\alpha S$ Lévy random field is used, and a reversible jump algorithm is implemented for making tractable Bayesian posterior inferences.

As oppose to the dimension reduction problem, chapter 3 focus on another set of problems where variables are either in or out of the model. Feature selection is done through the inference for the scale parameters in the diagonal Gaussian kernel functions. With a hierarchical mixture prior distribution of point mass at zero and a gamma distribution for the kernel scale parameters, the importance of each variable can be interpreted from the posterior inclusion probabilities. We introduced a $S\alpha S$ Lévy random measure for the kernel location parameters, which induces a mapping from the kernel function to a $S\alpha S$ random variable for modeling the unknown mean function. Continuous approximation for $S\alpha S$ Lévy random field is used, resulting a t_α prior distribution for the regression coefficients. Marginalizing the regression coefficients greatly improved the mixing of the Markov chain compared with the truncation method. We illustrate the Bayesian Additive Regression Kernels (BARK) model with various simulated and real data sets for both regression and classification problems.

Chapter 4 extends the BARK model to problems with a large number of vari-

ables. We looked at a particular set of problems where the variables exhibits an one dimensional Markov correlation structure. We present some preliminary results on the performance of kernel models with a large p for classification problems. Although the kernel models are efficient classifiers, they are sensitive to the choice of baseline, hence not so attractive for selecting out features.

Chapter 2

Bayesian Dimension Reduction with Kernel Logistic Models

2.1 Introduction

Advances in computing have broadened scientists' horizon for investigating information from many explanatory variables for classification problems. The response variable, however, may depend only on a small subset of the explanatory variables, or a few linear combinations of those variables. In situations with many predictor variables, it is common to use variable selection techniques (Clyde and George, 2004; George and McCulloch, 1997, 1993), or dimension reduction methods (Cook and Ni, 2005; Globerson and Tishby, 2003; Li, 1991) to reduce the model complexity.

While variable selection procedures are typically applied after specifying a particular model, dimension reduction methods can be used before identifying a specific parametric model, for example, see Chiaromonte and Cook (2002). With p explanatory variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, sufficient dimension reduction seeks a few independent linear combinations $U_1^T\mathbf{X}, \ldots, U_r^T\mathbf{X}$, such that all the information about the response Y is contained in these d linear combinations. The space spanned by $U_1^T, \ldots, U_r^T$ is called the dimension reduction space, and the goals are determining the smallest dimension d and the corresponding directions $U_1^T, \ldots, U_r^T$. In particular, if each U_l contains only one non-zero element, the dimension reduction space can be attained by variable selection procedures. Several methods have been proposed to find the reduction space. Principle Hessian directions (PHD; Li, 1992; Cook, 1998) locates the main axes along which the regression surface shows the largest curvature

in an aggregated sense; sliced inverse regression (SIR; Li, 1991; Duan and Li, 1991) looks at the first moment $E(\mathbf{X}|Y)$ in the inverse regression problem to learn about the dimension reduction space; sliced average variance estimation (SAVE; Cook and Weisberg, 1991) uses the first and the second moments in the inverse regression problem to estimate the directions. Recently, Wu *et al.* (2007) generalizes the SIR settings to non-linear dimension reduction using kernel models.

Given a set of directions $U_1, \ldots, U_l$, one may build a parametric model for classifying the response, and kernel models are one flexible family of parametric models. The regression model can be easily extended to classification problems through link functions. Kernel logistic regression (KLR; Cawley and Talbot, 2002, 2004) is commonly used for binary classification problems, which imposes a linear structure on the kernel functional space, capturing the complex non-linear relationship in the regression function. It differs from the Support Vector Machines (SVM; Boser *et al.*, 1992; Cristianini and Shawe-Taylor, 2000), because SVM focuses on estimating the optimal decision boundary separating difference classes, while KLR estimates the posterior probabilities of the class membership and subsequently establishes a decision boundary at some fixed threshold probability.

While KLR models are good classifiers for making predictions, it lacks the ability for interpreting the explanatory variables, especially in high dimension problems. In this chapter, we incorporate the dimension reduction idea into KLR through low rank Gaussian kernel functions. The Gaussian kernel is characterized by the precision matrix, which can be decomposed to a eigenvector matrix and corresponding eigenvalues. The $p \times r$ eigenvector matrix $\mathbf{U} = (U_1, \ldots, U_r)$ converts the original explanatory variables to d linear combinations of them $U_1^T\mathbf{X}, \ldots, U_r^T\mathbf{X}$, which spans the dimension reduction space. The dimension d can be determining through a series hypothesis testing (Shao *et al.*, 2007; Cook and Ni, 2005). In our nonparametric

KLR model, we preselect several candidates d, make a fully Bayesian inference for each d, and then compare those models with cross-validation error rates. The kernel functions are centered on the training samples like SVM, but the number of kernel functions is treated as a random variable. Furthermore, different eigenvalues in the Gaussian kernel precision matrix creates different kernel functions even if the kernel locations are the same. This leads to a flexible setting that could capture different type of interactions among the features in the dimension reduction space.

In the next section, we present the details of the nonparametric low rank kernel logistic regression model, which makes simultaneous inference for the kernel regression parameters and dimension reduction parameters. The prior distributions for the number of kernel functions, regression coefficients, and kernel parameters are jointly specified through an approximation for the $S\alpha S$ Lévy random field in Section 2.2.2. Lévy random field has been used in kernel regression problems (see Clyde *et al.*, 2005, 2006; Tu *et al.*, 2006; Clyde and Wolpert, 2007), and we extend it to KLR models for binary classification problems. The dimension reduction idea is implemented through the inference for the eigenvector matrix in the Gaussian kernels, which is further illustrated in Section 2.2.3. We developed a novel proposal distribution on the Stiefel manifold, which facilitates the MCMC for the dimension reduction projection matrix. Section 2.2.4 details the elicitation for the hyperparameters. Then we demonstrate the model with some simulation studies and real data analysis in Section 2.3. Finally, we conclude with a discussion in Section 2.4.

2.2 Method

In this section, we introduce the nonparametric kernel logistic regression model. We extend it to the nonparametric setting, and detail the joint prior specification for

the regression coefficients and the kernel parameters. We detail the choice of the prior distributions, which approximates the symmetric α stable process, and then we illustrate the elicitation for the hyperparameters. Finally, we present the algorithms to make posterior samples for all parameters.

2.2.1 Kernel Logistic Regression

Given a set of training samples

$$\{(y_i, \mathbf{x}_i)\}_{i=1}^n, \quad y_i \in \{0, 1\}, \quad \mathbf{x}_i \in \mathbb{R}^p,$$

where $\mathbf{x}_i$ is the vector of p explanatory variables for the ith training sample. Let $z_i = \mathsf{P}(y_i = 1 \mid \mathbf{x}_i) \in (0, 1)$ be the probability that the ith sample belongs to the class "1", then the logistic regression procedure aims to construct a linear model of the form

$$\text{logit}\{z\} = f(\mathbf{x}) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}, \qquad \text{where logit}(z) = \log\frac{z}{1-z}.$$

Assuming that the class labels y_is are drawn from independent distributed Bernoulli distributions with probabilities z_is, the likelihood of the data is given by

$$\prod_{i=1}^n z_i^{y_i}(1-z_i)^{1-y_i} = \prod_{i=1}^n \{g[f(\mathbf{x}_i)]\}^{y_i}\{1-g[f(\mathbf{x}_i)]\}^{1-y_i}, \tag{2.1}$$

where $g[f(\mathbf{x})] = 1/(1+\exp\{-f(\mathbf{x})\})$. This likelihood holds not only for the regular logistic function, but for more general regression functions $f(\mathbf{x})$. We shall mainly focus on the regression term $f(\mathbf{x})$ for the following kernel logistic regression models. A standard way to estimate the parameters for the regular logistic regression is through iterative weighted least squares algorithm.

A kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ induces a mapping from input vector $\mathbf{x}$ into a kernel function $K(\cdot, \mathbf{x})$, and KLR implements a linear logistic regression model on the kernel functional space,

$$\text{logit}\{z\} = f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i K(\mathbf{x}, \mathbf{x}_i).$$

Due to the non-linearity of the kernel mapping, this model is actually models the non-linear actions on the original covariate space. The kernel functions in KLR center at those n training sample points, and they share the same kernel parameters other than the location parameter. One way to estimate the parameters is via iterative algorithms, first estimate the regression coefficients by standard techniques conditional on the kernel parameters, by substituting the original variables by the kernel matrix, then estimate the kernel parameters by gradient method conditional on the regression coefficients, repeat those steps until convergence occurs.

In this chapter, we shall focus on rank d Gaussian kernels with following form

$$K_r(\mathbf{x}, \boldsymbol{\chi}, \boldsymbol{\Lambda}) \;=\; \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\chi})^T \boldsymbol{\Lambda} (\mathbf{x}-\boldsymbol{\chi})\right\}, \tag{2.2}$$

where the kernel precision matrix $\boldsymbol{\Lambda}$ is rank-d, *i.e.*

$$\boldsymbol{\Lambda} = \mathbf{U}\,\text{diag}\,\{\boldsymbol{\lambda}\}\mathbf{U}^T = \sum_{l=1}^{r} \lambda_l U_l U_l^T. \tag{2.3}$$

Here $\boldsymbol{\lambda} = \{\lambda_1, \cdots, \lambda_r\}$ is the eigenvalue vector of $\boldsymbol{\Lambda}$, and $\mathbf{U} = \{U_1, \cdots, U_r\}$ is the $p \times r$ eigenvector matrix of $\boldsymbol{\Lambda}$. Notice that $\lambda_l > 0$, and the column vectors in $\mathbf{U}$ are orthonormal, or $U_l^T U_{l'} = \delta_l^{l'}$. In other words, the eigenvector matrix $\mathbf{U}$ lies in the

Stiefel manifold

$$\mathcal{O}(p,r) \quad = \quad \{\mathbf{U} \in \mathbb{R}^{p\times r} \mid \mathbf{U}^T\mathbf{U} = \mathbf{I}_r\}. \tag{2.4}$$

In particular, when $r = p$, the Stiefel manifold becomes the orthogonal group of $p \times p$ orthogonal matrices.

With the matrix factorization (2.3), the kernel function (2.2) is also denoted as $K_r(\mathbf{x}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{U})$. Additional restrictions for the eigenvalue vector $\boldsymbol{\lambda}$ and eigenvector matrix $\mathbf{U}$ are needed for an identifiable model. The eigenvector matrix $\mathbf{U}$ serves as a dimension reduction matrix, which projects the point $\mathbf{x} \in \mathbb{R}^p$ to $\tilde{\mathbf{x}} = U^T\mathbf{x} \in \mathbb{R}^r$, and projects the point $\boldsymbol{\chi} \in \mathbb{R}^p$ to $\tilde{\boldsymbol{\chi}} = U^T\boldsymbol{\chi} \in \mathbb{R}^r$. In other words, the rank r Gaussian kernel function with parameter $\boldsymbol{\lambda}$ and $\mathbf{U}$ on the original covariate space can be treated as a full rank diagonal Gaussian kernel function on the dimension reduction space, *i.e.*

$$K_r(\mathbf{x}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{U}) = \exp\left\{-\frac{1}{2}(\tilde{\mathbf{x}} - \tilde{\boldsymbol{\chi}})^T \mathrm{diag}(\boldsymbol{\lambda})(\tilde{\mathbf{x}} - \tilde{\boldsymbol{\chi}})\right\} = K_r(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\chi}}, \boldsymbol{\lambda}, \mathbf{I}_r).$$

Dimension reduction techniques are commonly used to generate summary plot, which serve as a guidance to build parametric models. For example, when d is 1 or 2, it is possible to plot the projected data points on the dimension reduction space, and then discover important features that can be used for more sophisticated model building (Cook and Lee, 1999). Alternatively, we avoid any specific parametric model, and try to build a flexible nonparametric kernel regression model on the dimension reduction space.

2.2.2 nonparametric Kernel Logistic Regression

One disadvantage of the traditional KLR models is that all kernels have the same shape parameters, such as they share the common precision matrix $\mathbf{\Lambda}$ in the Gaussian kernel (2.2). Therefore, there are at most n different kernel functions in the regression term, where n is the number of training samples. Alternatively, if we allow some of the kernel shape parameters to vary for different kernel functions, the regression function could explore a much bigger functional space. In particular, with Gaussian kernel (2.2), suppose the regression function is a linear combination of random number of kernel functions,

$$f(\mathbf{x}) \quad = \quad \beta_0 + \sum_{j=1}^{J} \beta_j K_r(\mathbf{x}, \boldsymbol{\chi}_j, \boldsymbol{\lambda}_j, \mathbf{U}). \tag{2.5}$$

Under this formulation, all kernel functions share the same eigenvector matrix $\mathbf{U}$, but they are allowed to have different eigenvalues $\boldsymbol{\lambda}$ and location parameter $\boldsymbol{\chi}$. Denote by $\mathbf{X}$ the $n \times p$ covariate matrix, then the orthonormal matrix $\mathbf{U}$ projects the original covariate space to a dimension reduction space spanned by $\tilde{\mathbf{X}} = \mathbf{XU}$. The regression model (2.5) can be viewed as first project the original covariate space to a lower dimensional space, then build a kernel regression model on the lower dimensional space with diagonal kernel functions, whose eigenvalues do not need to be the same.

We need to construct a joint prior distribution for the projection orthonormal matrix $\mathbf{U}$, the number of kernels J, the kernel locations $\boldsymbol{\chi}_j$, the kernel eigenvalues $\boldsymbol{\lambda}_j$ and the regression coefficients β_0 and β_j, where $j = 1, \ldots, J$. Some of the parameters, such as β_0 and $\mathbf{U}$, do not need to change when the number of kernels changes, while other parameters, such as $\beta_{1:J}$, $\boldsymbol{\chi}_{1:J}$, and $\boldsymbol{\lambda}_{1:J}$, depend on the the number of kernels. They need to be treated differently.

Prior distributions for dimension changing parameters

Conditional on the number of kernel functions J, we would like to assign independent identical prior distributions for $(\beta_j, \boldsymbol{\chi}_j, \boldsymbol{\lambda}_j)$, $j = 1, \ldots, J$. Generally, suppose the prior for J is a Poisson distribution with an finite mean ν^+, the prior for β_j, $\boldsymbol{\chi}_j$, $\boldsymbol{\lambda}_j$ are independent from proper distributions with density function $\pi_\beta(\beta)$, $\pi_{\boldsymbol{\chi}}(\boldsymbol{\chi})$ and $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ respectively, it defines a Poisson random measure $\mathcal{N} \sim \mathsf{Po}(\nu)$, whose control measure is

$$\nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) = \nu^+ \pi_\beta(\beta)\pi_{\boldsymbol{\chi}}(\boldsymbol{\chi})\pi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}).$$

Denote by $\mathbb{X}$ the space the kernel location $\boldsymbol{\chi}$ lies, $\mathbb{R}^r_+ = \{(\lambda_1, \ldots, \lambda_r) \mid \lambda_1 > 0, \ldots, \lambda_r > 0\}$ the space kernel eigenvalue vector $\boldsymbol{\lambda}$ lies, then ν is a positive measure on $\mathbb{R} \times \mathbb{X} \times \mathbb{R}^r_+$. The kernel regression function (2.5) can be also represented as a stochastic integration of $\beta K(\cdot)$ with respect to this Poisson random measure, *i.e.*

$$f(\mathbf{x}) \quad = \quad \beta_0 + \int \beta K(\mathbf{x}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{U})\, \mathcal{N}(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}).$$

Rewrite the regression term as a stochastic integration of the kernel function K with respect to a signed measure

$$\mathcal{L}(d\boldsymbol{\chi}, d\boldsymbol{\lambda}) = \sum_j \beta_j \delta_{(\boldsymbol{\chi}_j, \boldsymbol{\lambda}_j)}(\boldsymbol{\chi}, \boldsymbol{\lambda}).$$

$\mathcal{L}$ can be viewed as a functional operator, who maps a L_2 functions $K(\mathbf{x}, \cdot, \cdot, \mathbf{U})$ to a random variable $\mathcal{L}[K(\mathbf{x}, \cdot, \cdot, \mathbf{U})] = \int K(\mathbf{x}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{U})\, d\,\mathcal{L}$, with characteristic function

$$\mathsf{E}\left[e^{it\mathcal{L}[K]}\right] \quad = \quad \exp\left\{\int \left(e^{it\beta K(\mathbf{x}, \boldsymbol{\chi}, \boldsymbol{\lambda}, \mathbf{U})} - 1\right) \nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda})\right\}.$$

Here ν is called the Lévy measure, the random signed measure $\mathcal{L}(d\boldsymbol{\chi}, d\boldsymbol{\lambda})$ is called the Lévy random measure, and the functional operator $\mathcal{L}(\cdot)$ is called the Lévy random field.

This procedure holds not only for finite measures ν, but also for more general positive measures. For example, the symmetric α-stable (SαS) Lévy random field takes Lévy measure

$$\nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) \quad = \quad \gamma c_\alpha |\beta|^{-1-\alpha} d\beta \pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi}) \pi_{\boldsymbol{\lambda}}(d\boldsymbol{\lambda}), \tag{2.6}$$

where $c_\alpha = (\alpha/\pi)\Gamma(\alpha)\sin(\pi\alpha/2)$, and $\pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi})$, $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ are probability measures on $\mathbb{X}$, $\mathbb{R}_+^r$ respectively. Here $0 < \alpha < 2$ is called the stable index, and $\gamma > 0$ is called the intensity parameter. Notice that the $S\alpha S$ Lévy random measure ν is not finite, directly sampling from the random variable $\mathcal{L}[K] = \int K \, d\mathcal{L}$ is not feasible. However, we can approximately sample from this random variable by truncating the regression coefficients $\boldsymbol{\beta}$. Omit all βs whose absolute magnitude is less than or equal to ϵ, where $0 < \epsilon < 1$ is some prespecified threshold, the stochastic integral $\mathcal{L}_\epsilon[K] = \int_{|\beta|>\epsilon} K \, d\mathcal{L}$ is well defined. We do not need to worry about the compensator functions, because the truncation is symmetric about 0, see appendix A for details.

The approximated SαS Lévy random field induces a joint prior distribution for the number of kernels J, the regression coefficients $\boldsymbol{\beta}$, kernel locations $\boldsymbol{\chi}$, kernel eigenvalues $\boldsymbol{\lambda}$,

$$\begin{aligned}
J &\sim \mathsf{Po}(\nu_\epsilon^+(\alpha, \gamma, \epsilon)) \\
\{\beta_j\}_{j=1}^J &\sim \frac{\alpha\epsilon^\alpha}{2} |\beta|^{-1-\alpha} 1_{[-\epsilon,\epsilon]^c}(\beta) d\beta \\
\{\boldsymbol{\chi}_j\}_{j=1}^J &\sim \pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi}) \\
\{\boldsymbol{\lambda}_j\}_{j=1}^J &\sim \pi_{\boldsymbol{\lambda}}(d\boldsymbol{\lambda})
\end{aligned}$$

where the expected number of kernel functions is

$$\nu_\epsilon^+ = \int_{[-\epsilon,\epsilon]^c \times \mathbb{X} \times \mathbb{R}^r} \nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) = \frac{2\gamma\Gamma(\alpha)}{\pi\epsilon^\alpha} \sin(\frac{\alpha\pi}{2}). \tag{2.7}$$

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

The regression coefficients associated with kernel functions have an identically independent two-sided Pareto prior distributions. The absolute value of these regression coefficient are restricted to be greater than ϵ, which is an acceptable assumption if ϵ is small. The tail of the Pareto distribution is polynomial, which is heavier than that of the normal distribution or the t distribution.

Like the most kernel regression models, we assume the kernel location parameters comes from the discrete set of the training samples. The kernel location space $\mathbb{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the collection of the training sample input vectors. It is natural to use a uniform prior distribution on this discrete set for the kernel locations, *i.e.* $\pi_{\boldsymbol{\chi}}(\boldsymbol{\chi}) = 1_{\mathbb{X}}(\boldsymbol{\chi})/n$.

Because the original p dimensional covariate space has already been reduced to a r dimensional subspace by the projection matrix $\mathbf{U}$, we would like to assume that all the eigenvalues in the rank d kernel are positive. Here we assign independent identically log normal prior distributions $\mathsf{LN}(\mu_\lambda, \sigma_\lambda^2)$ for the eigenvalues λ_l, $l = 1, \ldots, r$, whose density function is

$$\frac{1}{\sqrt{2\pi}\sigma_\lambda \lambda_l} \exp\left\{-\frac{(\log\lambda_l - \mu_\lambda)^2}{2\sigma_\lambda^2}\right\}.$$

Prior distributions for β_0 and $\mathbf{U}$

Since the intercept β_0 is always included in the kernel logistic regression model (2.5), we assign a double exponential prior distribution $\mathsf{DE}(b_0)$, whose density function is

$$b_0 \exp\{-b_0|\beta_0|\}/2.$$

Also $\mathbf{U}$ appears in every kernel function, a proper prior distribution on the Stiefel manifold (2.4) would be satisfactory. One natural choice is the uniform distribution on $\mathcal{O}(p, r)$. Given the rank d, all eigenvectors $\mathbf{U}$ lies in the same Stiefel manifold (2.4). We do not need to calculate the volume of $\mathcal{O}(p, r)$ in practice, because the ratio of the prior densities for two different $\mathbf{U}$s is 1.

2.2.3 Implementation

To make inference for the nonparametric logistic regression model, we need to draw samples of the parameters from the posterior distribution. The hyperparameters α, ϵ, a_λ, b_λ, b_0, μ_{μ_λ}, $\sigma^2_{\mu_\lambda}$, $a_{\sigma_\lambda^{-2}}$ and $b_{\sigma_\lambda^{-2}}$ are fixed by some default values. Because the data set have been standardized before the modeling, the default set up of the hyperparameters are not sensitive to the data sets.

The stable intensity γ has a Gamma hyper-prior distribution, which can be integrated out, and forms a negative binomial marginal prior distribution for the number of kernels. Other intermediate hyperparameters μ_λ, σ_λ^{-1} can be updated conjugate from their conditional distributions. Suppose the current model has J kernels, whose eigenvalue vectors are $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_J$, where $\boldsymbol{\lambda}_j = (\lambda_{j,1}, \ldots, \lambda_{j,r})$, then the conditional

distributions are

$$\mu_\lambda \mid \sigma^2_\lambda, \mu_{\mu_\lambda}, \sigma^2_{\mu_\lambda}, \boldsymbol{\lambda} \sim \mathsf{No}\left[\left(\sum_{j=1}^{J}\sum_{l=1}^{r}\log\lambda_{j,l} + \mu_{\mu_\lambda}\right)\Big/(rJ+1), \left(\frac{rJ}{\sigma^2_\lambda} + \frac{1}{\sigma^2_{\mu_\lambda}}\right)^{-1}\right],$$

$$\sigma^{-2}_\lambda \mid \mu_\lambda, \mu_{\mu_\lambda}, \sigma^2_{\mu_\lambda}, \boldsymbol{\lambda} \sim \mathsf{Ga}\left[a_{\sigma^{-2}_\lambda} + rJ/2, b_{\sigma^{-2}_\lambda} + \sum_{j=1}^{J}\sum_{l=1}^{r}(\log\lambda_{j,l} - \mu_\lambda)^2/2\right].$$

The regression intercept β_0 and the kernel eigenvector projection matrix $\mathbf{U}$ appears in every model. However, the likelihood (2.1) involves the logistic function, there is no conjugate update for these parameters, therefore we rely on Metropolis-Hasting steps to build the Monte Carlo Markov chain.

MCMC on the Stiefel Manifold

In order to make a random walk on the Stiefel manifold, we need to find a distribution with certain "center" and "spread" parameter on the Stiefel manifold. One choice is the von Mises-Fisher distribution (see Downs, 1972; Khatri and Mardia, 1977), which can be derived as the conditional distribution of a multivariate normal distribution on the unit sphere. Sampling from the von Mises-Fisher distribution can be done straightforwardly (see Hoff, 2007), but it is not very convenient to be used as a proposal distribution in the Metropolis-Hasting steps in our setting. On the other hand, projecting a multivariate normal distribution onto the unit sphere leads to another commonly used distribution for unit vectors. This is called the projected normal distribution, or the offset normal distribution by Mardia (1972). The projected normal distribution has a better interpretation in its parameters, which is also easier to sample from (see Presnell *et al.*, 1998; Nuñez Antonio and Gutiérrez-Peña, 2005).

Specifically, suppose X is a p dimensional multivariate normal random variable,

which has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, let U be the standardized X, *i.e.* $U = X/||X||$, where $||X|| = (X_1^2 + \cdots + X_p^2)^{1/2}$ is the L_2 norm. We call U has the projected normal distribution. Clearly the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are not identifiable without further constraints, since for any $c > 0$, taking $\boldsymbol{\mu}^* = c\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^* = c\boldsymbol{\Sigma}$ does not alter the distribution of the standardized variable. The difficulty could be address by restriction the determinant of $\boldsymbol{\Sigma}$ to be one. A common approach is to set $\boldsymbol{\Sigma} = \mathbf{I}_p$, the identity matrix, and the corresponding density function for U is (see Watson, 1983)

$$f(U) = \frac{\int_0^\infty x^{p-1}\phi(x - \boldsymbol{\mu}^T U)\, dx}{(2\pi)^{p/2} e^{||\boldsymbol{\mu}||^2/2}\phi(\boldsymbol{\mu}^T U)},$$

where ϕ and Φ are the standard normal density and cumulative distribution functions respectively.

Now we can construct the proposal distribution for the eigenvector matrix $\mathbf{U}$ from this vector form projected normal distribution. When $r < p$, let the proposal distribution for $\mathbf{U}$ only changes one column of the matrix at a time. First, select a column in $\mathbf{U}$ randomly, say U_l, and denote by S the vector space spanned by the remaining vectors in $\mathbf{U}$. Let H_S be the $p \times p$ projection matrix onto the space S, then $\mathbf{I}_p - H_S$ is the projection matrix onto the orthogonal space of S. Pre-select a number $\rho > 0$, generate a multivariate normal random variable $\mathbf{e} \sim \mathsf{No}(\mathbf{0}, I_p)$, project the vector $\rho U_l + \mathbf{e}$ to the orthogonal space of S, standardize it, and set it to $\tilde{U}_l$. In other words, $\tilde{U}_l$ is the following vector projected onto the unit sphere,

$$(\mathbf{I}_p - H_S)(\rho U_l + \mathbf{e}) = \rho U_l + (\mathbf{I}_p - H_S)\mathbf{e} \sim \mathsf{No}(\rho U_l, \mathbf{I}_p - H_S).$$

Notice that the vectors U_l and $\tilde{U}_l$ are orthogonal to column space S, therefore the probability density of obtaining $\tilde{U}_l$ from U_l from this proposal distribution is the same as the probability density of obtaining U_l from $\tilde{U}_l$. The proposal distribution

is symmetric, and we do not need to calculate the density in the Metropolis-Hasting steps.

When $r = p$, any matrix on the Stiefel manifold is rank p, We randomly select two columns U_{l_1} and U_{l_2}. First, ignore U_{l_2} and make a proposal for $\tilde{U_{l_1}}$ from U_{l_1} as if $r = p - 1$. Then project U_{l_2} to the orthogonal space spanned other vectors, including the newly proposed $\tilde{U_{l_1}}$. Similarly, the joint proposal distribution for U_{l_1} and U_{l_2} is also symmetric.

The scale parameter ρ in the proposal distribution controls the spread of the distribution. When $\rho = 0$, the proposal distribution for $\tilde{U_l}$ is a uniform distribution on $\mathcal{O}(p, r)$, and when $\rho = \infty$, the proposal distribution for $\tilde{U_l}$ is a point mass at U_l with probability one. By symmetry, the proposal density from U to U^* and from U^* to U is the same, therefore, the proposal density ratio $q(U \mid U^*)/q(U^* \mid U) = 1$.

Reversible Jump MCMC

The kernel regression coefficient β_j, kernel location parameter $\boldsymbol{\chi}_j$ and kernel eigenvalue vector $\boldsymbol{\lambda}_j$ depends on the total number of kernel functions J. Since they may involve dimension change in the model, these parameters are updated by the reversible jump MCMC algorithm (see Green, 1995).

We propose regular birth step from the prior distribution, and death step uniformly for existing kernels. In addition, when updating the regression coefficient β_j, we generate β_j^* from a t distribution centered at β_j. Since the β_j is restricted in $[-\epsilon, \epsilon]^c$, there is a positive probability that β_j^* lies within $[-\epsilon, \epsilon]$. In that case, we kill the corresponding kernel. When computing the Metropolis Hasting ratios, we need to account for the deaths from both the regular death step, and the out-of-domain update step.

To be more specific, write $\boldsymbol{\theta} = (\beta_{1:J}, \boldsymbol{\chi}_{1:J}, \boldsymbol{\lambda}_{1:J})$, and we omit other parameters

except J, because they are not changed in the reversible jump MCMC. With J kernels, suppose the probability to take a birth, death, or update step is $p_{b,J}$, $p_{d,J}$, or $p_{u,J}$ respectively, with $p_{b,J} + p_{d,J} + p_{u,J} = 1$. For a birth step, we propose $(J+1, \boldsymbol{\theta}^*)$ from $(J, \boldsymbol{\theta})$ by adding a new kernel $\theta^* = (\beta^*, \boldsymbol{\chi}^*, \boldsymbol{\lambda}^*)$. The probability to move $(J+1, \boldsymbol{\theta}^*)$ back to $(J, \boldsymbol{\theta})$ through a regular death step is $\frac{1}{J+1} p_{d,J+1}$, and through a out-of-domain update step is $\mathsf{P}(|\beta^*| < \epsilon) p_{u,J+1}$. Therefore, the acceptance rate for the birth step is the minimum of 1 and

$$\frac{p(\mathbf{y} \mid \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})} \frac{\frac{1}{J+1} p_{d,J+1} + \mathsf{P}(|\beta^*| < \epsilon) p_{u,J+1}}{\frac{1}{J+1} q(\theta^*) p_{b,J}},$$

where $\pi(\cdot)$ is the prior distribution density function, and $q(\cdot)$ is the the proposal distribution density function.

Similarly, the acceptance rate for $(J, \boldsymbol{\theta})$ from $(J+1, \boldsymbol{\theta}^*)$ through a regular death step is the minimum of 1 and

$$\frac{p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y} \mid \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)} \frac{q(\theta^*) p_{b,J}}{p_{d,J+1}},$$

and the acceptance rate for $(J, \boldsymbol{\theta})$ from $(J+1, \boldsymbol{\theta}^*)$ through a out-of-domain update step is the minimum of 1 and

$$\frac{p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y} \mid \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)} \frac{\frac{1}{J+1} q(\theta^*) p_{b,J}}{\mathsf{P}(|\beta^*| < \epsilon) p_{u,J+1}},$$

2.2.4 Elicit the hyperparameters

There are three hyper parameters α, γ, ϵ in this approximation of the SαS Lévy random field. The stable index α serves as the polynomial exponent in the two sided Pareto prior distribution for the regression coefficient β. We fix $\alpha = 1$, which

corresponds to the Cauchy process prior distribution.

The intensity parameter γ serves as a scalar in the expected number of kernels ν_ϵ^+ in (2.7). It measures the prior belief on how many kernels should be used in the regression function. We add a gamma hyper prior distribution $\mathsf{Ga}(a_\gamma, b_\gamma)$ for the intensity parameter γ. As a result, marginally the number of kernel functions J has a negative binomial prior distribution with size α and mean $2a_\gamma\Gamma(\alpha)\sin(\frac{\alpha\pi}{2})/(\pi b_\gamma\epsilon^\alpha)$. If we have some prior information on how many kernels should be used in the regression, for example, the probability no kernel is used, and the 95% quantile of the J, then we can solve the equations to elicit reasonable a_γ and b_γ.

The approximation parameter ϵ measures the minimal evidence for a kernel function to appear in the regression. It is typically determined by the computational power, and it is not very sensitive to the prediction result.

We have assigned independent identical prior distributions for the kernel eigenvalues, we need to assume that the projected features from the original covariate space are on the same scale. Since the projected variables are linear combinations of the original variables, whose weight vector lies on the unit sphere, the original variables should be on the same scale. Therefore, we standardize the variables before running the nonparametric kernel logistic regression. For the hyperparameters μ_λ and σ_λ^2, we assign normal inverse gamma conjugate hyper prior distributions

$$\mu_\lambda \sim \mathsf{No}(\mu_{\mu_\lambda}, \sigma^2_{\mu_\lambda}), \qquad \sigma_\lambda^{-2} \sim \mathsf{Ga}(a_{\sigma_\lambda^{-2}}, b_{\sigma_\lambda^{-2}}).$$

This allows us to combine the information of λ for different kernel functions. We fix $\mu_{\mu_\lambda}, \sigma^2_{\mu_\lambda}, a_{\sigma_\lambda^{-2}}, b_{\sigma_\lambda^{-2}}$ to allow the kernel eigenvalues to explore in a reasonable space for the standardized data, for example $\mu_{\mu_\lambda} = 0, \sigma^2_{\mu_\lambda} = 1, a_{\sigma_\lambda^{-2}} = 1, b_{\sigma_\lambda^{-2}} = 1$. This hyper prior distributions on μ_λ and σ_λ^2 collect information for the eigenvalues for all kernel

functions, and guide them towards high posterior probability regions.

2.3 Examples

We fit the low rank kernel logistic regression model for two simulated data sets and two real data sets. For simulation studies, we first generated 200 samples as training data set, then generated 1000 samples as testing data set, and recorded the out-of-sample prediction error rate. For real data analysis, we randomly split the data set into five parts, and train the model with four parts, and test it on the remaining part. We recorded the average out-of-sample prediction error rate in the 5-fold cross-validation study. Each experiment is repeated 20 times, and the performance of these models versus BART and SVM is shown in Figure 2.1. The variables of all studies are pre-standardized to mean 0 with standard deviation 1. We fit the low rank kernel regression model with $r = 1, 2$ and 3 for all data sets using the same default hyperparameters: $\alpha = 1, \epsilon = 0.1, a_\gamma = 1, b_\gamma = 1, \mu_{\mu_\lambda} = 0, \sigma^2_{\mu_\lambda} = 1, a_{\sigma_\lambda^{-2}} = 1, b_{\sigma_\lambda^{-2}} = 1, b_0 = 1$.

For the simulation studies, we can see that if the data is indeed generated from a lower dimensional structure, the low rank model will out perform SVM and BART. For the real data analysis, we do not know the if there is a low dimensional structure, the low rank model with small d still performs comparable with SVM and BART.

2.3.1 Simulation Studies

The first simulation study D4in1 have four variable, and the response variable is determined by one linear combination of those four variables. The variables $X_1, \ldots, X_4$ are

sampled from independent uniform distributions $\mathsf{Un}[-1,1]$, and the binary response

$$Y = \begin{cases} 1, & \text{if } X_1 + X_2 - X_3 - X_4 \in (-1,0] \cup (1,4]; \\ 0, & \text{if } X_1 + X_2 - X_3 - X_4 \in [-4,-1] \cup (0,1]. \end{cases}$$

The second simulation study D4in2 also have four variable, but the response variable is determined by two linear combinations of those four variables. Again, the variables $X_1, \ldots, X_4$ are sampled from independent uniform distributions $\mathsf{Un}[-1,1]$. The binary response

$$Y = \begin{cases} 1, & \text{if } (X_1 + X_2)(X_1 - X_3) \geq 0; \\ 0, & \text{if } (X_1 + X_2)(X_1 - X_3) < 0. \end{cases}$$

The out-of-sample prediction (Figure 2.1) shows that the rank 1 model is the best for D4in1 and the rand 2 model is the best for D4in2, and they are all better than SVM and BART. This is not surprising, because our simulated data are generated with a low dimensional structure in mind. In addition, the low rank model with a misspecified d does not perform as good as the one with the correct d. As a result, we could use the out-of-sample performance as a guidance to select the dimension d in practice.

Notice that although we are using the SAVE method to obtain the initial value of the eigenvector matrix in the low rank kernel, our method can adjust the directions automatically when the initial value via SAVE is not good. For example, in one simulation run of D4in1, SAVE estimated the direction as $0.63x_1 - 0.12x_2 - 0.09x_3 - 0.76x_4$, which hardly separates the two classes, see the left panel in Figure 2.2, while the posterior mean direction in the low rank 1 kernel regression model is $0.50x_1 + 0.51x_2 - 0.48x_3 - 0.51x_4$, which clearly separates the two classes, see the right panel

in Figure 2.2.

2.3.2 Real Data Analysis

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.

The out-of-sample prediction performance suggests the low rank model should choose $r = 2$ for Swiss bank data set, which is in line with the results in Cook and Lee (1999). Interestingly, SAVE detects three clusters along its first two directions, the low rank model with $r = 2$ detects another pattern which also well separates the genuine notes and the counterfeit notes, see Figure 2.3.

The Wisconsin diagnostic breast cancer (WDBC) data set (Wolberg *et al.*, 1995) contains 357 benign samples ($y = 0$) and 212 malignant samples ($y = 1$). The 30 real-valued variables are geometric features , such as radius, texture, smoothness of the cell nucleus. The task is to diagnose cancer from these geometric features.

Although the low rank models perform slightly worse than SVM for the WDBC data set, the directions obtained from the low rank models give us a new perspective to look at the data. Use the marginal tests for SAVE (Shao *et al.*, 2007), there is not a particular small dimension d that can present the full information for those 30 geometric features. If we plot the data on the first three SAVE directions, we could hardly see a separation for the cancer cases and the controls. However, with the rank 3 model, we are able to adjust the directions, and the plot on the posterior directions in the rank 3 model shows a much better separation for the cancer cases and the

controls, see Figure 2.4.

2.4 Discussion

In this chapter, we have described a Bayesian dimension reduction technique using kernel logistic models for classification problems. We introduced $S\alpha S$ Lévy random field to serve as the prior distribution on the unknown mean function, and the dimension reduction is done automatically through Bayesian posterior inference on the eigenvector matrix of the low rank Gaussian kernel functions. This approach can be extended to general regression problems where the response variable is continuous. Instead of using the logistic likelihood for the binary response, we can use normal or gamma likelihood for continuous response depending on specific regression problems. A fully Bayesian approach with low rank kernels will discover the low dimensional structure in the data, and the methodology still works.

As an alternative to dimension reduction, variable selection is wildly used to reduce the complexity of the model when the number of variables p is large. Instead of using a few linear combinations of the original variables, variable selection picks a subset of original variables, which is potentially more efficient when p is large. Lots of method have been proposed for Bayesian variable selection, for example, see George and McCulloch (1993, 1997); Clyde and George (2004). With a hierarchical prior distribution with a mixture of a point mass at 0 and a continuous distribution, we can make feature selections through the kernel scale parameters (see Chapter 3).

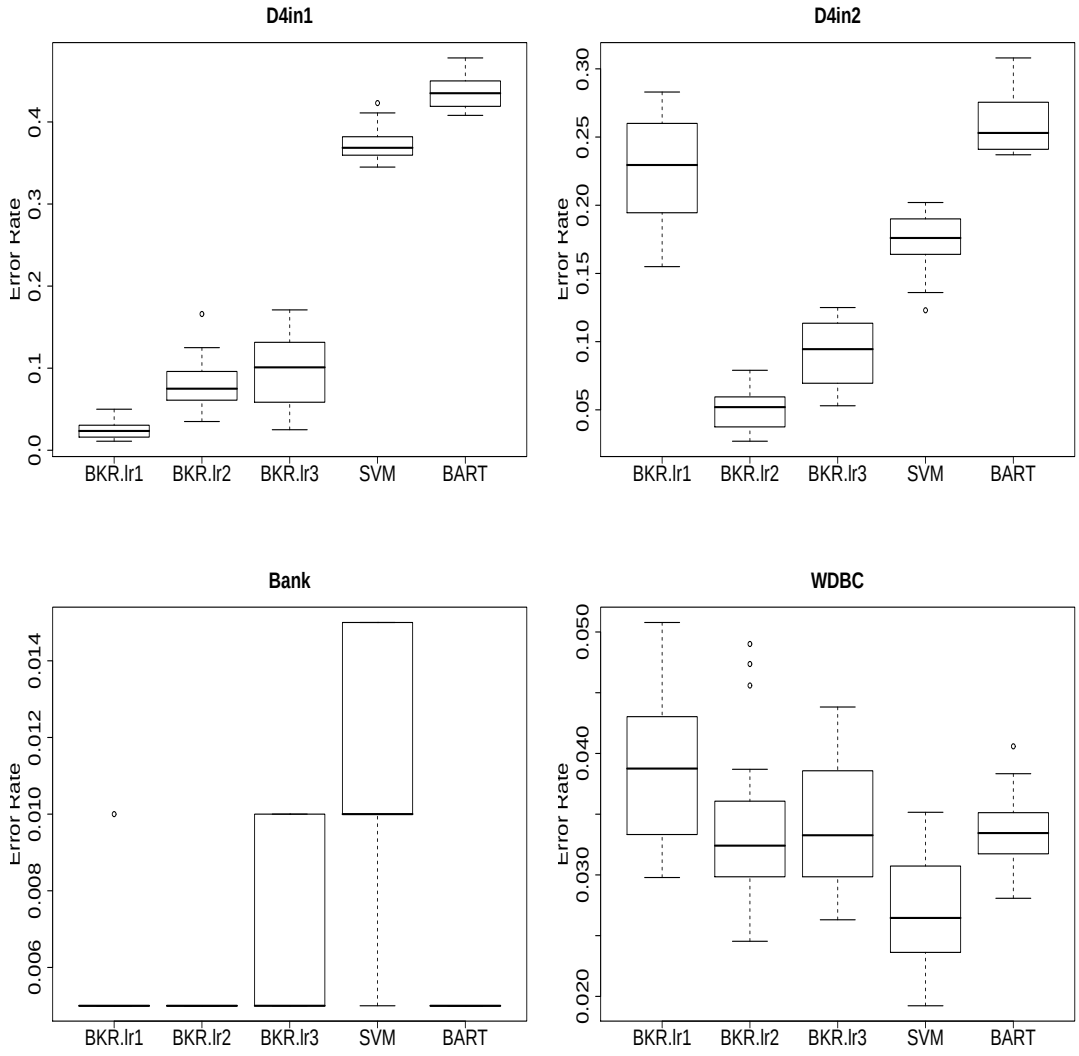

Figure 2.1: Performance comparison of the low rank model versus BART and SVM.

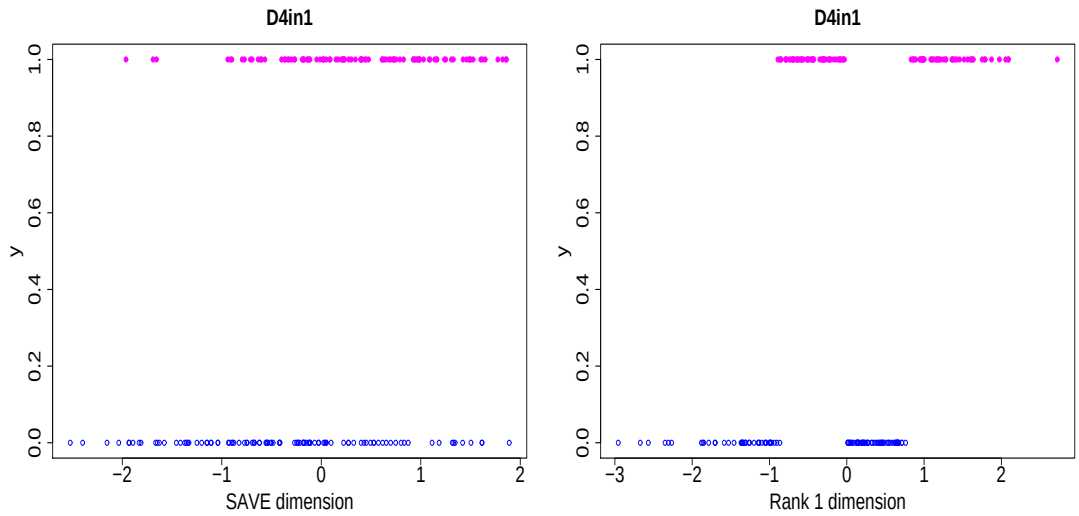

Figure 2.2: The feature dimension generated from SAVE (left) and the posterior mean of the rank 1 kernel model (right) in D4in1.

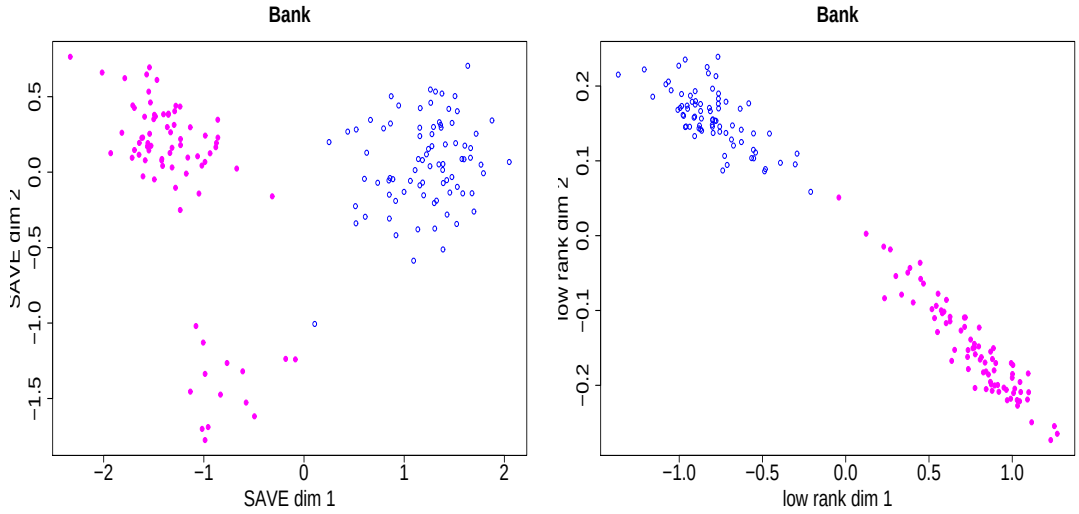

Figure 2.3: The feature dimension generated from SAVE (left) and the posterior mean of the rank 2 kernel model (right) in Swiss bank data set. The genuine notes are marked with blue open circles, and the counterfeit notes are marked with pink solid circles.

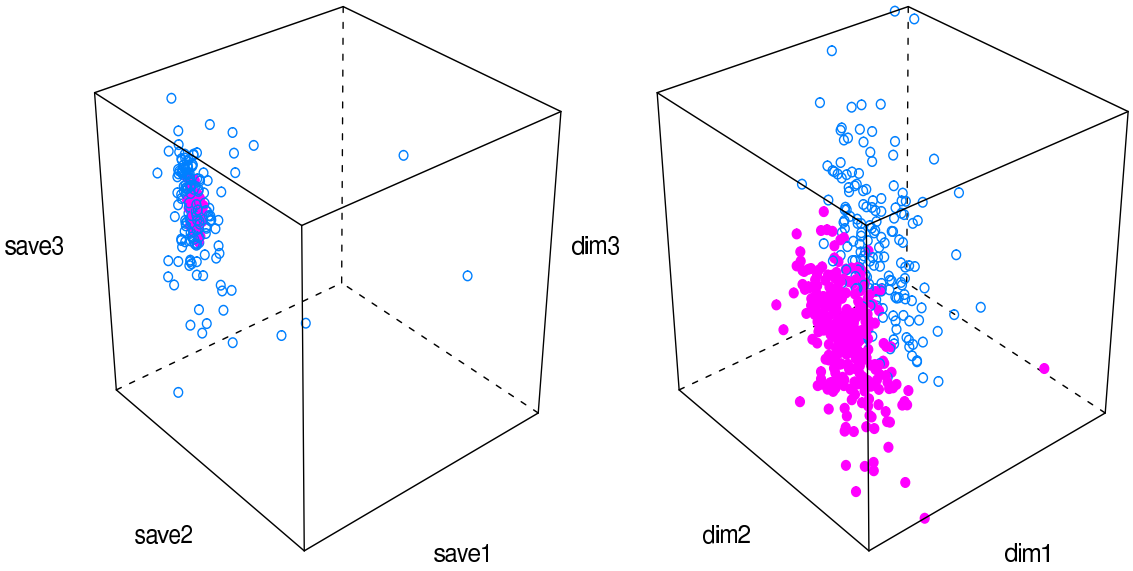

Figure 2.4: The feature dimension generated from SAVE (left) and the posterior mean of the rank 3 kernel model (right) in Wisconsin diagnostic breast cancer data set. The controls are marked with blue open circles, and the cancer cases are marked with pink solid circles.

Chapter 3

Bayesian Additive Regression Kernels with Feature Selection

3.1 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 multidimensional, say $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$. 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 $\mathsf{E}(Y \mid \mathbf{X} = \mathbf{x}) = f(\mathbf{x})$, and one popular candidate for the mean function is constructed by the sum of generating functions,

$$f(\mathbf{x}) \quad = \quad \sum_j g(\mathbf{x}, \boldsymbol{\theta}_j), \tag{3.1}$$

where $\boldsymbol{\theta}_j$ is the parameter in the jth 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 jth 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 (3.1) 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

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

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

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathsf{No}(0, \phi^{-1}). \tag{3.2}$$

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

$$\beta_j \overset{iid}{\sim} \mathsf{No}(0, \varphi_j^{-1}), \quad \varphi_j \overset{iid}{\sim} \mathsf{Ga}(a, b), \quad \phi \sim \mathsf{Ga}(c, d). \tag{3.3}$$

Both SVM and RVM search for the regression coefficient $\boldsymbol{\beta}$ that optimizes the target function. Since most of the β_js 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 (3.2), they used proper prior distribution for the full Bayesian analysis.

In this chapter, we detail another fully Bayesian framework for supervised learning with mean function (3.1), *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 (3.3). 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 chapter.

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 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 3.3. We demonstrate the approach through several simulated and real data sets in Section 3.4, and concludes in Section 3.5.

3.2 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

$$y_i = \sum_j \beta_j K(\mathbf{x}, \boldsymbol{\chi}_j) + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} \mathsf{No}(0, \phi^{-1}). \tag{3.4}$$

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,

$$f(\mathbf{x}) = \sum_j \beta_j K(\mathbf{x}, \boldsymbol{\chi}_j) = \iint_{\mathbb{R}\times\mathbb{X}} K(\mathbf{x}, \boldsymbol{\chi}) \mathcal{L}(d\boldsymbol{\chi}), \tag{3.5}$$

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 $\boldsymbol{\chi}$ with respect to probability measure $\pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi})$. 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.

3.2.1 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 (3.5) can be evaluated by drawing

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

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 (3.3) as b goes to zero and the number of kernels n goes to infinity. Denote by the $S\alpha S$ Lévy measure

$$\nu(d\beta, d\boldsymbol{\chi}) = \gamma c_\alpha |\beta|^{-1-\alpha} d\beta \pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi}), \tag{3.6}$$

where $c_\alpha = (\alpha/\pi)\Gamma(\alpha)\sin(\pi\alpha/2)$, and $\pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi})$ 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 $\mathcal{L}(A_j) \sim \mathsf{St}(\alpha, 0, \gamma\pi_{\boldsymbol{\chi}}(A_j), 0)$ (see Samorodnitsky and Taqqu, 1994, *pp.* 118).

In practice, Lévy random fields can be constructed from Poisson random measures, which can be further used in making posterior Bayesian inference, see 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 (3.6) 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 (3.5). One common approach is to truncate the mass β with respect to a given threshold $\epsilon > 0$. The Lévy measure is approximated by

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

which has a finite mass

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

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,

$$J \sim \mathsf{Po}(\nu_\epsilon^{T+}), \quad \{\beta_j\}_{1\le j\le J} \overset{iid}{\sim} \frac{\alpha\epsilon^\alpha}{2} |\beta|^{-\alpha-1} 1_{|\beta|>\epsilon}\, d\beta, \quad \{\boldsymbol{\chi}_j\}_{1\le j\le J} \overset{iid}{\sim} \pi(\boldsymbol{\chi}), \tag{3.7}$$

where the prior distribution for the regression coefficients β's are called two-sided Pareto distributions. The approximated Lévy random Field $\mathcal{L}_\epsilon^T$ 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:

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

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

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

which has a finite mass

$$\nu_\epsilon^{C+}(\alpha,\gamma,\epsilon) = \frac{\gamma\alpha^{1-\alpha/2}}{2^{1-\alpha}\epsilon^\alpha}\frac{\Gamma(\alpha/2)}{\Gamma(1-\alpha/2)}. \tag{3.8}$$

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 $\boldsymbol{\beta}$ and kernel locations $\boldsymbol{\chi}$,

$$J \sim \mathsf{Po}(\nu_\epsilon^{C+}(\alpha,\gamma,\epsilon)), \quad \{\beta_j\}_{1\le j\le J} \overset{iid}{\sim} t(\alpha, 0, \epsilon^2), \quad \{\boldsymbol{\chi}_j\}_{1\le j\le J} \overset{iid}{\sim} \pi(\boldsymbol{\chi}), \tag{3.9}$$

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

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

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 (3.5) 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* $\nu(d\beta, d\boldsymbol{\chi}, du) = \gamma c_\alpha |\beta|^{-1-\alpha} d\beta\, \pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi})\, du$ *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* $\pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi})$ *is a probability measure on* $\mathbb{X}$*. It induces a Lévy random field* $\mathcal{L}$ *that maps a function* $g \in L_2(\mathbb{X}, \pi_{\boldsymbol{\chi}}(d\boldsymbol{\chi}))$ *to the random variable*

$$\begin{aligned} \mathcal{L}[g] &= \int_{\mathbb{R}\times\mathbb{X}\times(0,1)} (\beta - \sin\beta) g(\boldsymbol{\chi}) \mathcal{N}(d\beta, d\boldsymbol{\chi}, du) + \\ &\quad \int_{\mathbb{R}\times\mathbb{X}\times(0,1)} \sin\beta\, g(\boldsymbol{\chi}) \tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}, du) \end{aligned} \tag{3.10}$$

where

$$\mathcal{N} \sim \mathsf{Po}(\nu), \quad \tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}, du) = \mathcal{N}(d\beta, d\boldsymbol{\chi}, du) - \nu(d\beta, d\boldsymbol{\chi}, du).$$

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

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

$$\mathcal{L}_\epsilon^C[g] = \int_{\mathbb{R}\times\mathbb{X}\times(0,1)} \beta g(\boldsymbol{\chi}) 1_{\left\{u < (1+\alpha\epsilon^2\beta^{-2})^{-(\alpha+1)/2}\right\}}(u) \mathcal{N}(d\beta, d\boldsymbol{\chi}, du). \tag{3.11}$$

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

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

$$\mathsf{E}\left|\mathcal{L}[g] - \mathcal{L}_\epsilon^C[g]\right|^2 \leq ||g||_2^2 \frac{2\gamma}{\pi} \Gamma(\alpha+1) \sin\left(\frac{\pi\alpha}{2}\right) \left(\frac{(1+\alpha)\alpha^{\alpha/2}}{2} + \frac{\alpha^{(2-\alpha)/2}}{2-\alpha}\right) \epsilon^{2-\alpha} \tag{3.12}$$

In particular, for the Cauchy the squared discrepancy can be calculated exactly when

$\alpha = 1$,

$$\mathsf{E}\Big|\mathcal{L}[g] - \mathcal{L}_\epsilon^C[g]\Big|^2 = \int_{\mathbb{R}\times\mathbb{X}} g(\boldsymbol{\chi})^2 \frac{\gamma}{\pi}\left(1 - \frac{1}{\beta^2+\epsilon^2}\right) d\beta d\chi = \gamma\epsilon||g||_2^2. \tag{3.13}$$

This offers guidance on the choice of parameters γ and ϵ, which is further discussed in Section 3.2.4.

3.2.2 Sparse Representation

In this section, we shall detail the remaining prior distributions for BARK (3.4) 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.* $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. 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 $\mathbb{X} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n\}$, 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

$$\{\boldsymbol{\chi}_j\}_{1\le j\le J} \overset{iid}{\sim} \mathsf{Un}(\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}). \tag{3.14}$$

This prior specification guarantees that there are at most $n+1$ distinct kernels in

the model (3.4). When ϵ goes to zero, the prior distribution induces more and more kernels in the expression (3.4), 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

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

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 (3.9-3.14) becomes

$$\begin{aligned} J &\sim \mathsf{Po}(\nu_\epsilon^+(\alpha, \gamma, \epsilon)) \\ \mathbf{n} \mid J &\sim \mathsf{MN}(J, \mathbf{1}/(n+1)) \\ \{\tilde{\beta}_i\}_{0 \le i \le n} \mid \mathbf{n} &\overset{ind}{\sim} \mathsf{Cauchy}(0, n_i^2 \epsilon^2) \end{aligned} \tag{3.15}$$

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

3.2.3 Feature Selection

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

$$K(x, \boldsymbol{\chi}) = \exp\left\{ -\sum_{l=1}^{p} \lambda_l (x_l - \chi_l)^2 \right\}, \tag{3.16}$$

where the scale parameters λ_ls measure the contribution of the lth 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 lth variable through the kernel function; on the other hand, if λ_l is large, the lth 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 (3.16) 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.

BARK with equal weights

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

$$\lambda_l = \lambda, \quad \lambda \sim \mathsf{Ga}(a_\lambda, p b_\lambda),$$

where $l = 1, \ldots, p$, and p is the total number of variables. The exponent term in the Gaussian kernel (3.16) 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 3.4 falls into this category.

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_ls is $\mathsf{Ga}(a_\lambda, b_\lambda)$. Notice that gamma distribution is infinitely divisible, we can splits the total gamma mass equally into individual scale parameters λ_l. Specifically,

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

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 3.4 falls into this category.

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

$$\lambda_l = \begin{cases} \lambda_l^*, & \text{if } \delta_l = 1 \\ 0, & \text{if } \delta_l = 0 \end{cases}, \quad \{\delta_l\}_{1\le l\le p} \overset{iid}{\sim} \mathsf{Bi}(1, p_\lambda), \quad p_\lambda \sim \mathsf{Be}(a_p, b_p). \tag{3.17}$$

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 (3.17), let

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

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 3.4 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 date set (Johnson, 1995) in Section 3.4. falls into this category.

BARK with selection and different weights

Similarly, allowing the non-zero scale parameters to be different yields the fourth setting. Specifically, on top of (3.17), let

$$\lambda_l^* \overset{iid}{\sim} \mathsf{Ga}(a_\lambda/d, b_\lambda), \tag{3.18}$$

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.

3.2.4 Elicitation

There are three parameters in the SαS 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 $\boldsymbol{\beta}$. 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 γ and ϵ as follows. The approximation threshold ϵ serves as the scale parameter in the Cauchy prior distribution for the regression coefficient β, and both ϵ and the intensity parameter γ determines the number of kernels J in BARK. With 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 (3.13). In particular, for kernel function $K(\mathbf{x}, \boldsymbol{\chi})$ with a discrete uniform prior for $\boldsymbol{\chi}$ on $n+1$ locations, the L_2 discrepancy

$$\mathsf{E}|\mathcal{L}[K(\mathbf{x}, \cdot)] - \mathcal{L}_\epsilon[K(\mathbf{x}, \cdot]|^2 = \gamma\epsilon||K(\mathbf{x}, \cdot)||_2^2 = \frac{\gamma\epsilon}{n+1}\sum_{i=0}^{n} K(\mathbf{x}, \boldsymbol{\chi}_i)^2$$

With a smaller ϵ, 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 γ and ϵ through observable quantities the expected number of kernels γ/ϵ 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 (3.16) 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 avoid, 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 3.2.3, 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 $(x_l - \chi_l)^2/2 \sim \mathsf{Ga}(1/2, 1/2)$, 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 interquantile 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 $\mathsf{E}K$ is in $[0.3, 0.5]$, which can be used to elicit the intensity parameter γ.

In BARK with selection, the probability that each λ_l is non-zero has a Beta prior

distribution in (3.17). 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

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

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 $\mathsf{P}(d = k) = \frac{1}{p+1}$ for $k = 0, 1, \ldots, p$. In other words, the expected number of signal variables in the prior specification is $k/2$. One can also elicit the hyperparameters 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 $\mathsf{Ga}(c, d)$ 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 (3.4).

3.2.5 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 $\mathbf{y} = \{y_1, \ldots, y_n\}^T$ 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

$$p(y^* \mid \mathbf{y}) = \int p(y^* \mid \mathbf{x}^*, \phi, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda}) p(\phi, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\lambda} \mid \mathbf{y}) \, d\phi \, dJ \, d\boldsymbol{\beta} \, d\boldsymbol{\chi} \, d\boldsymbol{\lambda} \tag{3.19}$$

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 SαS Lévy random field (3.7), the regression coefficients have independent symmetric Pareto prior distributions with density

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

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_α prior distributions (3.15) directly on the regression coefficients directly in the continuous approximation. However, we can represent t_α 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

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

Conditional on the number of kernels J and kernel locations $\{\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_J\}$, 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 $\mathbf{i} = (i_1, \ldots, i_m)$, *i.e.* $n_{i_j} > 0$ for $j = 1, \ldots, m$. 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 $1 \le j \le n$, $0 \le k \le n$. Denote by K^* the $n \times m$ sub-matrix of $\tilde{K}$, where $K^*_{j,k} = K_{j,i_k}$.

After integrating out $\boldsymbol{\beta}^*$, the likelihood is

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

where

$$\boldsymbol{\Sigma}^* = \left(\phi K^{*T} K^* + \operatorname{diag}\left\{\frac{\varphi_1^*}{n_{i_1}^2}, \ldots, \frac{\varphi_m^*}{n_{i_m}^2}\right\}\right)^{-1}, \quad \boldsymbol{\mu}^* = \phi \boldsymbol{\Sigma}^* K^{*T} \mathbf{y}. \tag{3.20}$$

Denote $\boldsymbol{\theta} = (\mathbf{n}, \boldsymbol{\varphi}^*, \boldsymbol{\lambda}, \phi)$, after integrating out $\boldsymbol{\beta}$, 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 $\boldsymbol{\beta}^*$ is $\mathsf{No}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ is defined in (3.20). 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 $\frac{1}{M}\sum_{m=1}^{M} f(\mathbf{x} \mid \boldsymbol{\theta}^{(m)}, \boldsymbol{\beta}^{*(m)})$ with $\boldsymbol{\beta}^{*(m)} \sim \mathsf{No}(\boldsymbol{\mu}^{*(m)}, \boldsymbol{\Sigma}^{*(m)})$ 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, details see appendix C. By integrating out $\boldsymbol{\beta}$, 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.

3.3 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 chapter, we focus on binary classification, where the response variable $y \in \{0, 1\}$. With the Probit link function,

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

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

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

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 $\boldsymbol{\beta}$ 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 (3.20), 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 $\mathsf{No}(f(\mathbf{x}_i), 1)$ truncated above zero; if $y_i = 0$, the conditional distribution for z_i is $\mathsf{No}(f(\mathbf{x}_i), 1)$ 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.

3.4 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.

3.4.1 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 3.1. We demonstrate the performance of our model by three simulation studies and three real data sets.

Table 3.1: Predictive mean square errors in regression problems.

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

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

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

The corresponding regression mean functions are

$$\begin{aligned} f_2(\mathbf{x}) &= (x_1^2 + (x_2x_3 - 1/(x_2x_4))^2)^{1/2}, \\ f_3(\mathbf{x}) &= \arctan((x_2x_3 - 1/(x_2x_4))/x_1). \end{aligned}$$

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 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 3.1 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 λs on ZN, CHAS and B are tiny, hence we conclude that these three variables does not effect the housing price much.

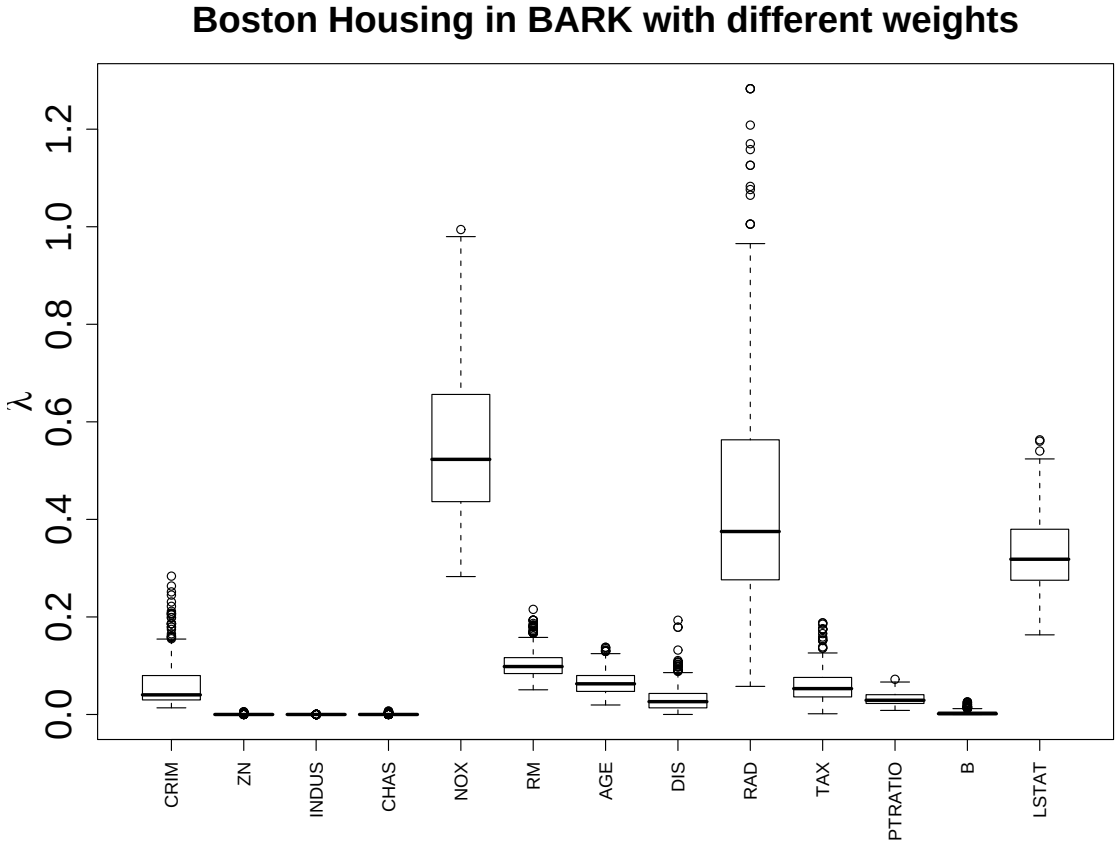

Figure 3.1: Box plots for the kernel scale parameters in Boston Housing data set in BARK with different weights.

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 3.1 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 3.2.

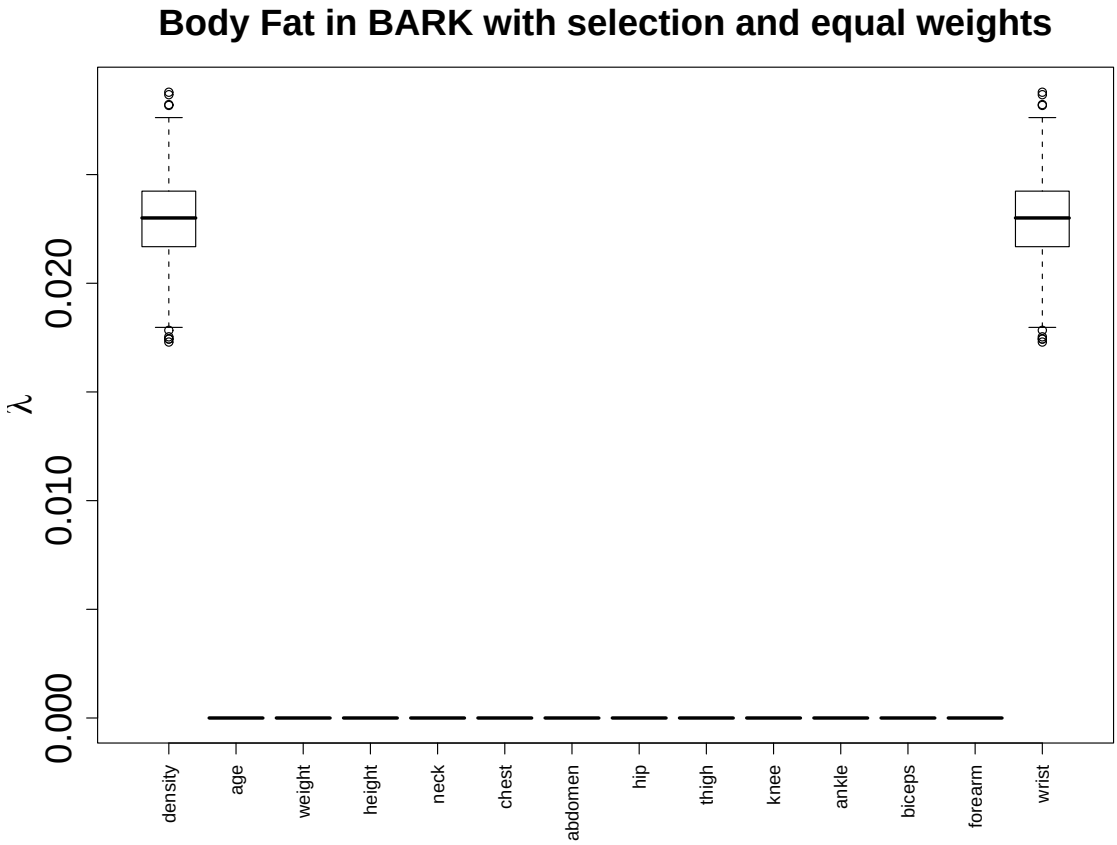

Figure 3.2: Box plots for the kernel scale parameters in body fat data set in BARK with selection and equal weights.

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 3.1 shows that the BARK models, SVM and BART are comparable in terms of out-of-sample prediction mean square errors.

3.4.2 Classification Examples

For classification problems, we calculate the predictive misclassification rate, and report the average of the 20 replicated runs in table 3.2. 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, $y = 1_{\{x_1^2+x_2^2 \leq 2/\pi\}}(\mathbf{x})$. Under this formulation, both class have roughly the same number of sample points.

Table 3.2: Predictive mis-classification rate in classification problems.

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 %

As we can see from Table 3.2, 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) 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 3.2 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 3.3 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. In Chapter 2, we extends BARK with lower rank models to capture those structures.

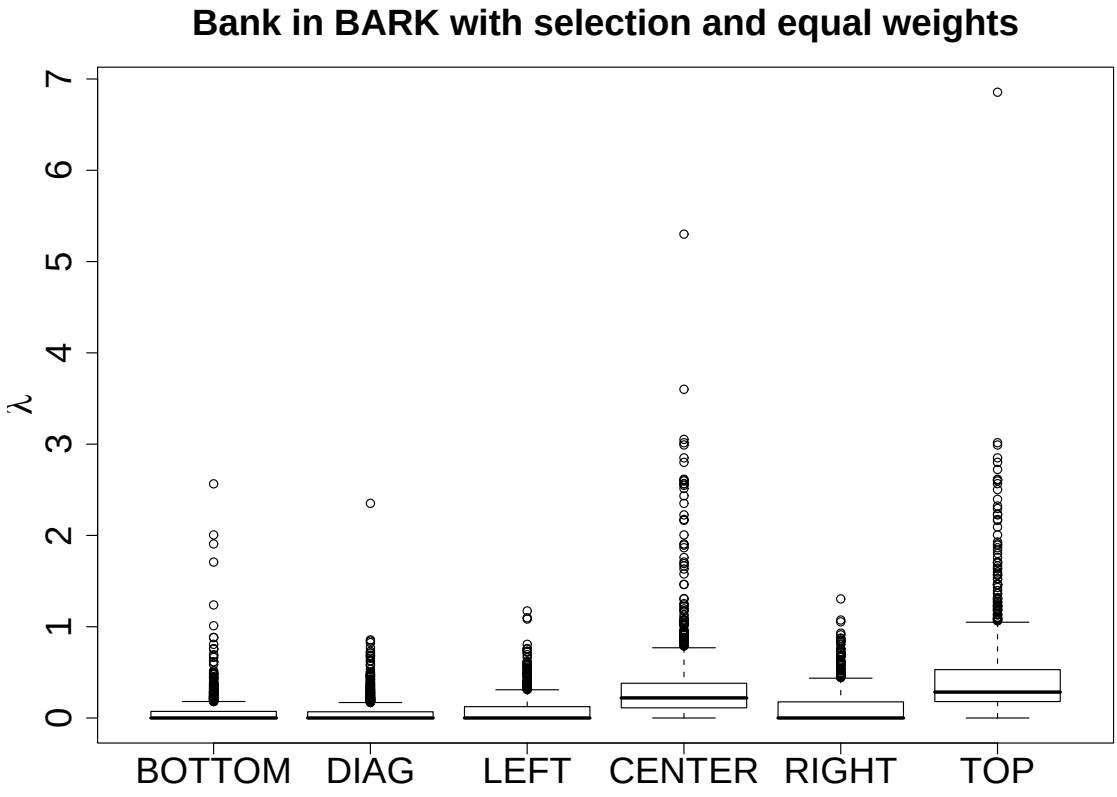

Figure 3.3: Box plots for the kernel scale parameters in Swiss bank note data set in BARK with selection and equal weights.

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 3.2 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 3.2 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.

3.5 Discussion

In this chapter, 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 3.2 admits a number of extensions. In this chapter, 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 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 3.3, 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 3.3 is applicable for the multi-class classification problems.

Chapter 4

Structural Baysian Additive Classification Kernels

4.1 Introduction

Variable and feature selection has become the focus of scientific research when hundreds or thousands of variables are available. This is very common in biomedical applications, such as analyzing protein mass spectroscopy and single nucleotide polymorphism (SNP) data sets. The scientific goals are to identify regions of interest among the explanatory variables that differentiates samples from different experimental groups. We got inspired by a Matrix-assisted laser desorption/ionization (MALDI) data set in the breast cancer study. Consider a classification problems with n observations, the response variables $y_1, y_2, \ldots, y_n$ are either 0 or 1, indicating samples are in the normal tissue group or in the invasive tumor tissue group. For each sample i, we observe the intensity x_{ij} for many time-of-flight values t_j, where $j = 1, 2, \ldots, p$. Time of flight values are directly associate to the mass over charge ratio m/z of the corresponding proteins in the tissue samples. In this study, we looked at 24 normal tissue samples and 56 invasive tumor tissue samples. The goal is to find m/z regions that discriminate between two groups. A unique feature of this data set is that the variables are ordered a priori according to the m/z values. In other words, if a peak is observed at an m/z site, the intensities of nearby m/z sites are high. Similar structure can be found in SNP data, where the variables are aligned according the biology order on the DNA.

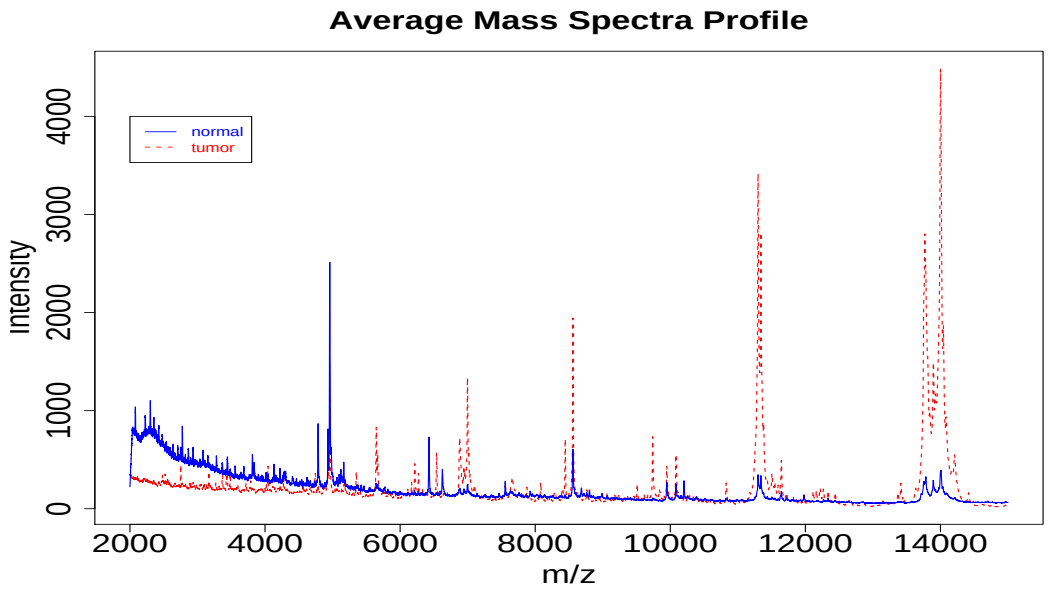

Figure 4.1: Average mass spectroscopy data for normal tissues (blue solid line) and invasive tumor tissues (red dash line).

Tibshirani *et al.* (2005) introduced the fused Lasso model to tackle this problem. It penalizes the L_1 norm of both the regression coefficients and their successive differences, encouraging sparsity of the coefficients and also sparsity of their differences, *i.e.* local constancy of the coefficient profile. This shed light to Bayesian treatment of this problem by putting correlated structures on the prior distribution for the regression coefficients. Recently, Li and Zhang (2008) proposed a Bayesian variable selection method in structured high-dimensional covariate space using Ising prior distributions on the regression coefficients, and we would like to extend this idea to kernel regression models.

The kernel function maps the original covariate space to a functional feature space, hence the regression coefficients measure the contributions of each kernel function in stead of the original variables. Selecting few kernel functions in the regression through the regression coefficient corresponds to a sparse representation, while selecting some

non-zero kernel scale parameters corresponds to selections on the original variables, which is easier to interpret. For example, in a diagonal Gaussian kernel function,

$$K(\mathbf{x}, \boldsymbol{\chi}, \boldsymbol{\lambda}) = \exp\left\{-\lambda_1(x_1 - \chi_1)^2 - \lambda_2(x_2 - \chi_2)^2 - \cdots - \lambda_p(x_p - \chi_p)^2\right\}$$

the kernel scale parameter λ_l is the bridge linking variable X_l to the regression function. Shrinking λ_l to zero effectively drops out the variable X_l in the regression. Chapter 3 described a feature selection procedure through a hierarchical prior distribution on λ_l's in the Bayesian Additive Regression Kernels (BARK) model. However, this approach cannot to used directly in large p problems, because exploring all possible combinations of the variables are very expensive in the *one-change-at-a-time* MCMC scheme. Because the non-linearity from the kernel functions joins the Gaussian likelihood for the regression mean function, direct sampling from the posterior distribution is impossible, and the data augmentation trick in the Swendsen-Wang algorithm (Higdon, 1998) cannot be applied.

To overcome this difficulty, we propose a novel proposal distributing that accommodate the correlated structural prior distribution on the kernel scale parameters. In the next section, we detail the Markov prior structure, and the relationship to the commonly used Ising prior distribution. We demonstrate how to use the novel proposal distribution to make Metropolis-Hasting updates. And we illustrate the elicitation of the hyper-parameters for the Markov prior distribution. Next, we demonstrate the performance and limitations of this model through some simulated data set and real mass spectroscopy data set in Section 4.3. Finally, we end with a discussion on potential extensions and future direction in Section 4.4.

4.2 Bayesian Additive Classification Kernels with Dependence Prior Structure

Continue with the Bayesian Additive Classification Kernels (BACK) model in Section 3.3, the model can be summarized by

$$\begin{aligned} y_i &= 1(z_i > 0) \\ z_i &= \sum_{j=1}^{J} \beta_j K(\mathbf{x}_i, \boldsymbol{\chi}_j, \boldsymbol{\lambda}) + \epsilon_i \\ \epsilon_i &\sim \mathsf{No}(0,1) \\ J &\sim \mathsf{Po}\left(\nu_\epsilon^{C+}\right) \\ \{\beta_j\}_{1\le j\le J} \mid J &\overset{iid}{\sim} t(\alpha, 0, \epsilon^2) \\ \{\boldsymbol{\chi}_j\}_{1\le j\le J} \mid J &\overset{iid}{\sim} \mathsf{Un}\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\} \end{aligned}$$

where ν_ϵ^{C+} is the mass of the continuous approximated $S\alpha S$ Lévy measure in (3.8). All kernel functions share the same scale parameter $\boldsymbol{\lambda}$. Since the number of variables p is large compared with the number of observations n, there is not enough information to discriminate different contributions for all variables. Similar to prior specification in BARK with selection and equal weights in Section 3.2.3, we restrict all the non-zero scale parameters have the the same value. Specifically, denote $\delta_l = 1(\lambda_l > 0$, conditional on the indicator vector $\boldsymbol{\delta}$,

$$\lambda_l \mid \delta_l = \begin{cases} \lambda, & \text{if } \delta_l = 1 \\ 0, & \text{if } \delta_l = 0 \end{cases} \qquad \lambda \mid \boldsymbol{\delta} \sim \mathsf{Ga}(a_\lambda, d\, b_\lambda)$$

where $d = \sum_{l=1}^{p} \delta_l$ is the number of non-zero kernel scale parameters in $\boldsymbol{\lambda}$. Instead of using independent identically Bernoulli prior distribution on δ_l as in Section 3.2.3, we introduce the correlation on the scale parameters from the following prior distri-

bution.

4.2.1 Structural Prior Specification

The prior distribution on the indicator vector $\boldsymbol{\delta} \in \{0,1\}^p$ is specified through a Markov model, which is uniquely determined by an initial Bernoulli prior distribution on δ_1 and a transition matrix. Specifically, let $P(\delta_1 = 1) = q_1^{(init)}$, and the transition matrix

$$A = \begin{pmatrix} q_0 & 1-q_0 \\ 1-q_1 & q_1 \end{pmatrix} \tag{4.1}$$

i.e., $P(\delta_{l+1} = 0 \mid \delta_l = 0) = q_0$ and $P(\delta_{l+1} = 1 \mid \delta_l = 1) = q_1$. Therefore, the stationary distribution of this Markov chain has a Bernoulli distribution with success rate $q = (1-q_0)/(2-q_0-q_1)$. It is typical to set the initial probability $q_1^{(init)}$ to be the stationary probability q, in which case the distribution for δ_1 and δ_p at both ends of the indicator vector are close when p is large.

Suppose there are d 1's in $\boldsymbol{\delta}$, which are coming from k continuous blocks of 1's, then we can calculate the prior probability for $\boldsymbol{\delta}$ explicitly. Denote by

$$A(q_0, q_1) = q_0^{p-d-k} q_1^{d-k} (1-q_0)^k (1-q_1)^k / (2-q_0-q_1),$$

1. If $\delta_1 = 0$, $\delta_p = 0$, there are $k+1$ blocks of 0's, and

$$P(\boldsymbol{\delta}) = (1-q) q_0^{p-d-(k+1)} q_1^{d-k} (1-q_0)^k (1-q_1)^k = \frac{1-q_1}{q_0} A(q_0, q_1)$$

2. If $\delta_1 = 0$, $\delta_p = 1$, there are k blocks of 0's, and

$$P(\boldsymbol{\delta}) = (1-q)q_0^{p-d-k}q_1^{d-k}(1-q_0)^k(1-q_1)^{k-1} = A(q_0, q_1)$$

3. If $\delta_1 = 1$, $\delta_p = 0$, there are k blocks of 0's, and

$$P(\boldsymbol{\delta}) = (1-q)q_0^{p-d-k}q_1^{d-k}(1-q_0)^{k-1}(1-q_1)^k = A(q_0, q_1)$$

4. If $\delta_1 = 1$, $\delta_p = 1$, there are $k-1$ blocks of 0's, and

$$P(\boldsymbol{\delta}) = (1-q)q_0^{p-d-(k-1)}q_1^{d-k}(1-q_0)^{k-1}(1-q_1)^{k-1} = \frac{q_0}{1-q_1}A(q_0, q_1)$$

Let $B(\boldsymbol{\delta}) = (q_0/(1-q_1))^{\delta_1+\delta_p-1}$ be the deviation effect of the boundary terms, then the prior probability mass can be summarized as $P(\boldsymbol{\delta}) = B(\boldsymbol{\delta})A(q_0, q_1)$.

On the other hand, Ising prior distribution is commonly used on 0/1 vectors. The prior probability for $\boldsymbol{\delta}$ is defined up to a normalizing constant

$$\pi_{\boldsymbol{\delta}}(\boldsymbol{\delta}) = Z(\alpha_\delta, \beta_\delta)^{-1} \exp\left(\alpha_\delta \sum \delta_l + \beta_\delta \sum \delta_l \delta_{l+1}\right),$$

where the normalizing constant $Z(\alpha_\delta, \beta_\delta)$ is typically very difficult to calculate due to the complex combinatorics for large p. The benefit of using Ising prior distribution is that the full conditional distribution for one particular location is only depend on its adjacent locations, and does not involve the normalizing constant, *i.e.*

$$\mathsf{P}(\delta_l = 1 \mid \boldsymbol{\delta}_{-l}) = \mathsf{P}(\delta_l = 1 \mid \delta_{l-1}, \delta_{l+1}) = \frac{\exp[\,\alpha_\delta + \beta_\delta(\delta_{l-1} + \delta_{l+1})]}{1 + \exp[\,\alpha_\delta + \beta_\delta(\delta_{l-1} + \delta_{l+1})]}, \qquad (4.2)$$

for $l = 1, 2, \ldots, p$ and set $\delta_0 = \delta_{p+1} = 0$ as the boundary condition. This is also

called the auto-logistic model by Besag (1972, 1974). There is a close relationship between the Markov prior and the Ising prior. In fact, if there are d 1's in $\boldsymbol{\delta}$, which are coming from k continuous blocks of 1's, the prior probability for $\boldsymbol{\delta}$ under the Ising prior distribution is

$$P(\boldsymbol{\delta}) \quad = \quad Z(\alpha_\delta, \beta_\delta)^{-1} \exp\left\{\alpha_\delta d + \beta_\delta (d-k)\right\},$$

which is very similar to the probabilities from the Markov prior distribution. Set

$$\alpha_\delta = \log\left[\frac{(1-q_0)(1-q_1)}{q_0^2}\right], \quad \beta_\delta = \log\left[\frac{q_0 q_1}{(1-q_0)(1-q_1)}\right].$$

It is straightforward to verify that the full conditional probabilities from the Markov prior distribution is also (4.2) for $l = 2, 3, \ldots, p-1$. which only differs on the boundaries. From simulation, the difference of two specifications are not big when p is large. We advocate the Markov prior specification, because it can be easily extend to making inference for different hyper-parameters q_0 and q_1, since the normalizing constant can be evaluated explicitly when calculating the prior probability ratio for different hyper-parameters.

4.2.2 Novel Proposal Distributions

Conditional on the $S\alpha S$ parameters $\{\alpha, \gamma, \epsilon\}$ and the kernel regression parameters $\{\mathbf{z}, J, \boldsymbol{\beta}, \boldsymbol{\chi}, \lambda\}$, the posterior distribution for the indicator vector $\boldsymbol{\delta}$ is proportional to

$$\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(z_i - \sum_{j=1}^{J}\beta_j K\left(\mathbf{x}_i, \boldsymbol{\chi}_j, \lambda \mathrm{diag}(\boldsymbol{\delta})\right)\right)^2\right\} \times \frac{(db_\lambda)^{a_\lambda}}{\Gamma(a_\lambda)}\lambda^{a_\lambda - 1}e^{-db_\lambda \lambda} \times \pi_{\boldsymbol{\delta}}(\boldsymbol{\delta}) \quad (4.3)$$

where $d = \sum_{l=1}^{p} \delta_l$ is the number of non-zero δ's in the indicator vector, and $\pi_{\boldsymbol{\delta}}(\boldsymbol{\delta})$ is the prior probability mass function for $\boldsymbol{\delta}$ as described in Section 4.2.1. The posterior probability for $\boldsymbol{\delta} = \mathbf{0}$ is slightly different, because the kernel is then the intercept, and the kernel scale parameter λ vanished. We need to use reversible jump algorithm for this particular case, and regular Metropolis-Hasting algorithm is applicable when both $\boldsymbol{\delta}$ and the proposed new $\boldsymbol{\delta}^*$ are not vector of all zeros.

The simplest way to make a move for $\boldsymbol{\delta}$ is randomly flip the value at a particular location. Imagine a sparse problem, which expects most variables are not relevant. In other words, $\boldsymbol{\delta}$ have much more zeros than ones. A flip procedure at a completely random location may not be efficient, because it flips a zero to one most of the time under this scenario. Alternatively, we may prefer having similar probability to flip a one to zero and flip z zero to one. Treat the consecutive 1's in $\boldsymbol{\delta}$ a meaningful block unit in the structural high dimensional problem, a flip move may corresponds to five different type of "logical moves" in terms of the block units.

1. Birth of a new block. For example, $\boldsymbol{\delta} = (\cdots, 0, 0, 0, \cdots)$, $\boldsymbol{\delta}^* = (\cdots, 0, 1, 0, \cdots)$, then the prior probability ratio for $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}$ is

$$\frac{\mathsf{P}(\boldsymbol{\delta}^*)}{\mathsf{P}(\boldsymbol{\delta})} = \frac{B(\boldsymbol{\delta}^*)}{B(\boldsymbol{\delta})} \times \frac{(1-q_0)(1-q_1)}{q_0^2}.$$

2. Death of a existing block of single 1. For example, $\boldsymbol{\delta} = (\cdots, 0, 1, 0, \cdots)$, $\boldsymbol{\delta}^* = (\cdots, 0, 0, 0, \cdots)$, then the prior probability ratio for $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}$ is

$$\frac{\mathsf{P}(\boldsymbol{\delta}^*)}{\mathsf{P}(\boldsymbol{\delta})} = \frac{B(\boldsymbol{\delta}^*)}{B(\boldsymbol{\delta})} \times \frac{q_0^2}{(1-q_0)(1-q_1)}.$$

3. Merge of two very close blocks into one block. For examples, $\boldsymbol{\delta} = (\cdots, 1, 0, 1, \cdots)$,

$\boldsymbol{\delta}^* = (\cdots, 1, 1, 1, \cdots)$, then the prior probability ratio for $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}$ is

$$\frac{\mathsf{P}(\boldsymbol{\delta}^*)}{\mathsf{P}(\boldsymbol{\delta})} = \frac{B(\boldsymbol{\delta}^*)}{B(\boldsymbol{\delta})} \times \frac{q_1^2}{(1-q_0)(1-q_1)}.$$

4. Split of a block into two very close blocks. For example, $\boldsymbol{\delta} = (\cdots, 1, 1, 1, \cdots)$, $\boldsymbol{\delta}^* = (\cdots, 1, 0, 1, \cdots)$, then the prior probability ratio for $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}$ is

$$\frac{\mathsf{P}(\boldsymbol{\delta}^*)}{\mathsf{P}(\boldsymbol{\delta})} = \frac{B(\boldsymbol{\delta}^*)}{B(\boldsymbol{\delta})} \times \frac{(1-q_0)(1-q_1)}{q_1^2}.$$

5. Stretch or recess a block of 1's one location on the left or the right end. This does not change the total number of blocks, and we plan to incorporate it to more general Stretch and recess moves.

If we decide to make a flip move, first calculate all possible locations that will yield a birth, a death, a merge or a split move. We postpone the stretch and recess moves to more general moves that keep the number of blocks unchanged. Then we pick a possible type of move at random, and finally pick a random location of that type of move to finish the proposal flip. For example, if we move $\boldsymbol{\delta}$ to $\boldsymbol{\delta}^*$ via a birth flip from this proposal distribution, then the proposal probability ratio is

$$\frac{q(\boldsymbol{\delta} \mid \boldsymbol{\delta}^*)}{q(\boldsymbol{\delta}^* \mid \boldsymbol{\delta})} = \frac{P(\text{a death flip location} \mid \text{death move in } \boldsymbol{\delta}^*)\, P(\text{death move} \mid \boldsymbol{\delta}^*)}{P(\text{a birth flip location} \mid \text{birth move in } \boldsymbol{\delta})\, P(\text{birth move} \mid \boldsymbol{\delta})}.$$

Combining this with the posterior probability (4.3), we can calculate the Hasting's ratio in the Metropolis-Hasting algorithm easily.

Keeping the number of blocks unchanged, we can also propose moves that change $\boldsymbol{\delta}$ at multiple locations. Three natural moves are

1. Stretch one end of a block with k additional 1's, while avoiding combining two

blocks together. For example, move $\boldsymbol{\delta} = (*, 1, 0, ..., 0, 0, *)$ to $\boldsymbol{\delta}^* = (*, 1, 1, ..., 1, 0, *)$, or move $\boldsymbol{\delta} = (*, 0, 0, ..., 0, 1, *)$ to $\boldsymbol{\delta}^* = (*, 0, 1, ..., 1, 1, *)$. Therefore, the prior probability ratio for $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}$ is

$$\frac{\mathsf{P}(\boldsymbol{\delta}^*)}{\mathsf{P}(\boldsymbol{\delta})} = \frac{B(\boldsymbol{\delta}^*)}{B(\boldsymbol{\delta})} \times \frac{q_1^k}{q_0^k}.$$

2. Recess one end of a block to delete k 1's, while keeping the block have at least one 1. For example, move $\boldsymbol{\delta} = (*, 1, 1, ..., 1, 0, *)$ to $\boldsymbol{\delta}^* = (*, 1, 0, ..., 0, 0, *)$, or move $\boldsymbol{\delta} = (*, 0, 1, ..., 1, 1, *)$ to $\boldsymbol{\delta}^* = (*, 0, 0, ..., 0, 1, *)$. Then the prior probability ratio for $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}$ is
$$\frac{\mathsf{P}(\boldsymbol{\delta}^*)}{\mathsf{P}(\boldsymbol{\delta})} = \frac{B(\boldsymbol{\delta}^*)}{B(\boldsymbol{\delta})} \times \frac{q_0^k}{q_1^k}.$$

3. Shift one block of 1s to the left (or right), while keeping both the total number of 1's and the total number of blocks unchanged (avoiding the merging of two sections). It is easy to see the prior probability ratio reduced to $\mathsf{P}(\boldsymbol{\delta}^*)/\mathsf{P}(\boldsymbol{\delta}) = B(\boldsymbol{\delta}^*)/B(\boldsymbol{\delta})$.

The general proposal scheme that keep the number of blocks unchanged is described as follows. First, calculate possible types of moves for all ends of each block in $\boldsymbol{\delta}$, and the the maximum step of each type of move. Then propose the step size of the move according to a prespecified distribution. For example, we used a discritized truncated normal density whose standard deviation depends on the length of block, such at longer blocks can be stretched/recessed with more locations. For example, if we move $\boldsymbol{\delta}$ to $\boldsymbol{\delta}^*$ via a stretch move of step size k on the left end of the first block (denoted as $(1, L)$) from this proposal distribution, then the proposal probability

ratio is

$$\frac{q(\boldsymbol{\delta} \mid \boldsymbol{\delta}^*)}{q(\boldsymbol{\delta}^* \mid \boldsymbol{\delta})} = \frac{P(\text{recess with size } k \mid \text{recess on } (1, L) \text{ of } \boldsymbol{\delta}^*)\, P(\text{recess on } (1, L) \mid \boldsymbol{\delta}^*)}{P(\text{stretch with size } k \mid \text{stretch on } (1, L) \text{ of } \boldsymbol{\delta})\, P(\text{stretch on } (1, L) \mid \boldsymbol{\delta})}.$$

4.2.3 Elicitation

The elicitation for the parameters in the $S\alpha S$ prior distribution is described in Section 3.2.4, and we will only focus on the elicitation for q_0 and q_1 in the Markov prior distribution for the indicator vector $\boldsymbol{\delta}$. Let $q = (1-q_0)/(2-q_0-q_1)$ be the probability of 1 in the stationary distribution of the Markov chain, Li and Zhang (2008) recommends using $r = (1-q_0)/(1-q_1) = q/(1-q)$ and $w = q_0/(1-q_1)$ to parametrize the model, where r is the prior odds of $\delta_l = 1$, and w reflects the increase in probability of $\delta_l = 0$ if we know $\delta_{l-1} = 0$. To our experience, we found it more convenient to elicit the prior distribution through q and $s = 1 - q_1$. Assuming that there are p variables in total, pq is the expected number of 1's in $\boldsymbol{\delta}$, and pqs is the expected number of blocks of 1's in $\boldsymbol{\delta}$. In other words, $1/s$ is the expected length of 1 blocks.

All pairs of $(q_0, q_1) \in (0,1)^2$ yield valid prior specification, but not all pairs of $(q, s) \in (0,1)^2$ are legitimate. In fact, $sq/(1-q) = 1-q_0 \in (0,1)$, hence $s < (1-q)/q$. In other words, we can select q and s from the prior information on how many variables and blocks of variables should be involved in the regression, and transform them back to q_0 and q_1 to specify the Markov prior distribution on $\boldsymbol{\delta}$ via

$$q_0 = 1 - \frac{qs}{1-q}, \qquad q_1 = 1 - s.$$

For example, if we expect approximately 100 variables in 10 blocks are relevant among 1000 variables, set $q = s = 0.1$, *i.e.* $q_0 = 0.989$, $q_1 = 0.9$.

4.3 Examples

We first illustrate this method using a very nice simulated data set. Suppose there are 5 simulated spectra in group $y = 0$ and 5 simulated spectra in group $y = 1$. Each spectra have three peaks with 100 mass to charge ratio values: one is unique in the each group, one one shares the same location with same intensities in two groups, and the remaining one shares the same location but has different intensities in two groups. The simulated data is shown in Figure 4.2, where spectra with $y = 1$ are plotted in blue solid lines, and spectra with $y = 0$ are plotted in red dash lines. To avoid overlaying lines, we plot the red lines downwards, but with the same intensity scale.

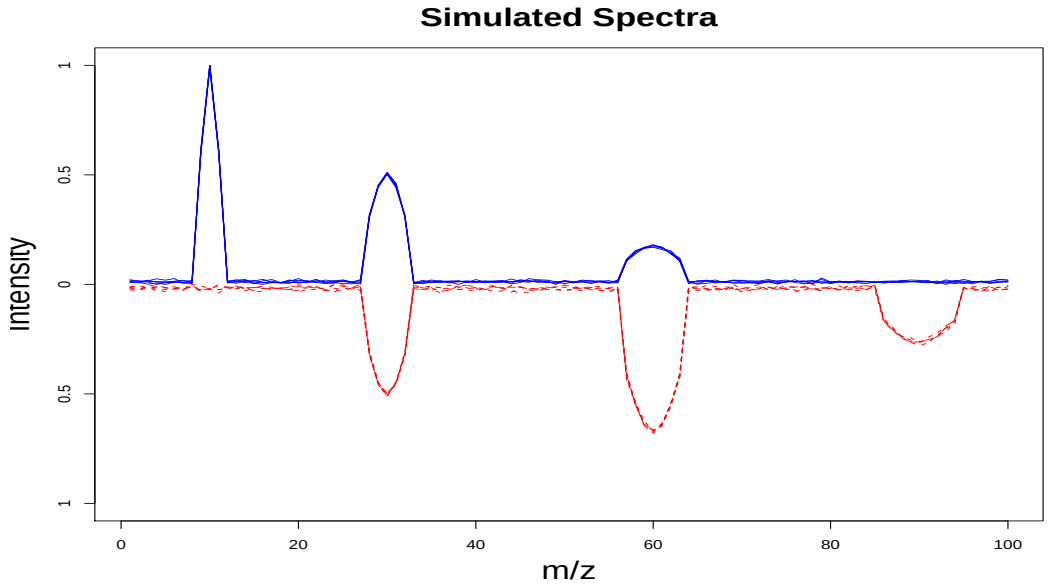

Figure 4.2: Simulated 3-peak mass spectra with length 100, 5 with response $y = 0$ (blue solid lines) and 5 with response $y = 1$ (red dash lines).

In this simulated study, we expect two groups differ in about 20 out of 100 mass to charge locations, hence we set $q = 0.2$. In addition, we expect to see three block of different sections, hence we set $s = 0.15$. Figure 4.3 shows the posterior inclu-

sion probabilities of each $\lambda_l > 0$ with different starting values of $\boldsymbol{\delta}$. We have tried four different starting values for $\boldsymbol{\delta}$, all zeros (denoted as 00), all ones (denoted as 11), alternative zeros and ones (denoted as 01, and a estimate from the quantile of the difference of the mean spectrum in two groups (denoted as *diff*). After $100,000$ burning iterations, we kept 1000 iterations out of the following $100,000$ iterations for posterior inference (*i.e.* one out of every 100 samples are kept). The inclusion probability Figure 4.3 captures the locations that the intensities are differently expressed in two groups, and the mixing of the chain is good, since the curve from different starting values are very similar.

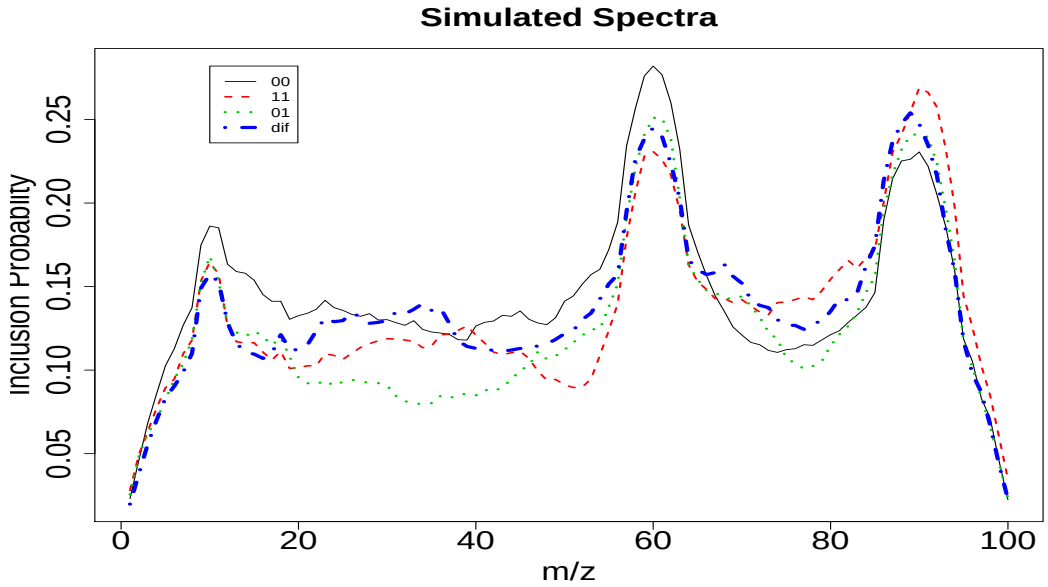

Figure 4.3: Inclusion probabilities for the kernel scale parameters λ_l's in BACK, with $q = 0.2$ and $s = .15$.

The inclusion probability plot does reveal three peaks, which corresponds to the three locations that the intensities are expressed differently in two groups. In addition, it does not pick up the location of the common peak with same intensity in two groups, which is expected because the kernel functions in BACK only uses the

difference of the explanatory curve and the kernel center curve. However, we see a relative high baseline at locations that two groups have the similar intensities in the inclusion probability curve. This is due to the Markov prior structure, which makes adjacent locations correlated. Furthermore, we notice that the peak in the inclusion provability curve is not particularly high, only about 0.3 in this simulation study, which is much less than 1. The result is not satisfactory, but is reasonable in BACK. The likelihood only interested in the classification problem, which only want to find some regions that completes the classification task. In the simulated spectra, even some consecutive mass-to-charge locations have different intensities in two groups, maybe just one of the location (for example, the center of the peak location) can already complete the classification task. The Markov prior structure pushes the adjacent locations into the model, which have similar effect as the middle location and create redundancy in the model. Similarly, maybe one peak is good enough to discriminate two groups, so there is also redundancy in multiple peaks in this model. Like variable selection in linear model, if there are two co-linear covariates, the posterior model will pick each variable half of the time, the inclusion provability curve demonstrate this *dilution* in BACK.

We also looked at a small section of the mass spectroscopy data in the breast cancer study, where the mass to charge ratio is ranging from 6391 Dalton to 6464 Dalton, shown in Figure 4.4. The baseline for each mass spectrum is different. A typical multistage approach to analyze mass spectroscopy data involves pre-processing, which removes the baseline and normalizes each spectrum, for example, see Noy and Fasulo (2007); Morris *et al.* (2005); Tibshirani *et al.* (2004). We use locally-weighted polynomial regression (LOWESS Cleveland and Devlin, 1988; Cleveland, 1979) to estimate the baseline for each mass spectrum. Notice that the peak width increases as mass to charge ratio increases, we apply LOWESS on the intensity versus the nat-

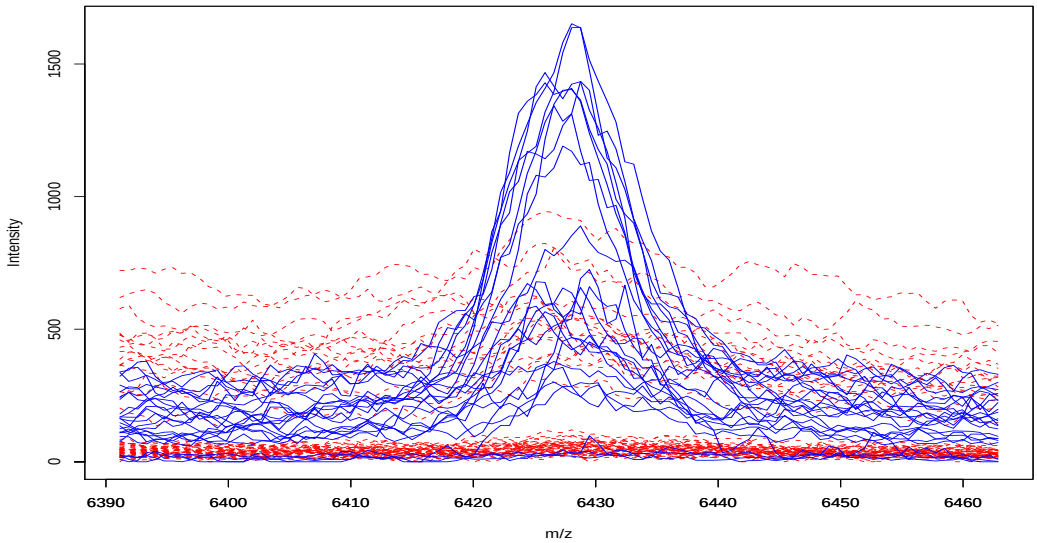

Figure 4.4: A small section of the mass spectroscopy data for normal tissues (blue solid line) and invasive tumor tissues (red dash line).

ural logarithm of the mass to charge ratio. Figure 4.5 shows the normalized spectra after subtracting the LOWESS baseline (denoted as baseline 1), along with the posterior inclusion probability plot for each mass-to-charge location. Clearly something is wrong, as the inclusion probability plot does not recover the "obvious" different peak in the normal group and the invasive tumor group. A closer look at the difference of the intensities for the mean spectra in two group reveals that the difference at all locations are greater than zero, see the black curve in Figure 4.7. This suggests that any location is a good indicator to discriminant two groups. If we additionally subtract a small constant for the mass spectra of the normal tissues, pulling the difference of the two group mean spectra towards zero, see the pink curve in Figure 4.7. With this new baseline subtraction (denoted by baseline 2) in Figure 4.6, the inclusion probability detects the difference peak for discriminating two groups. From this really simple real data analysis, we find that the baseline subtraction is crucial

in interpreting BACK model.

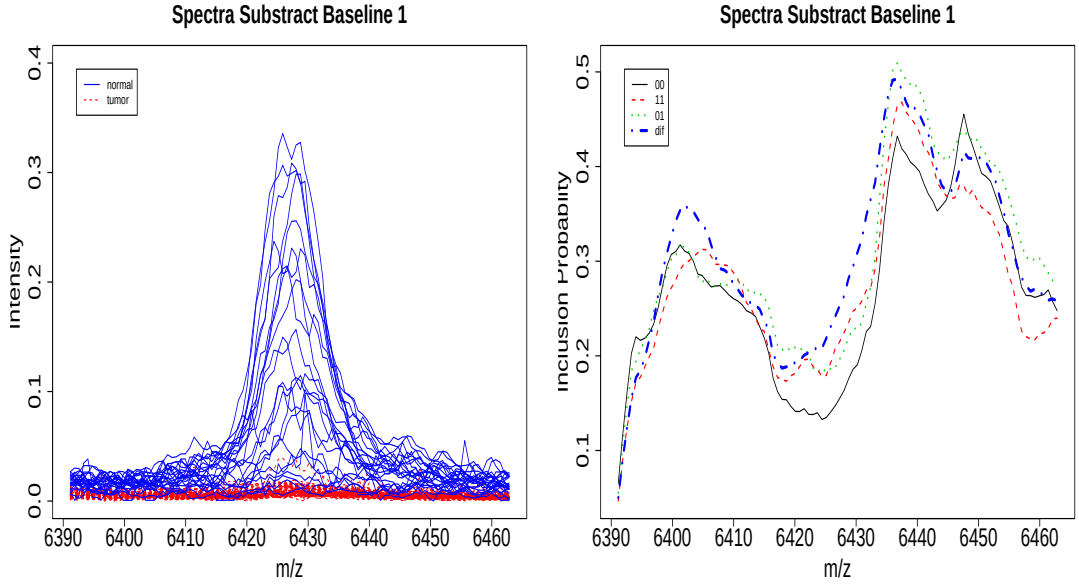

Figure 4.5: Spectra subtracting baseline 1 and the inclusion probabilities.

4.4 Discussion

In this chapter, we explored the Bayesian Additive Classification Kernels with a structural prior distribution on kernel parameters when the number of variables p are large. We illustrated this method using fixed hyper-parameters in the Markov prior distribution (or equivalently the Ising prior distribution for large p) in some simulated examples and real data analysis. We found this method have two major drawbacks: dilution on the inclusion probability from both adjacent locations and multiple blocks, and sensitivity to the initial choice of baseline. This warns us although kernel methods are very effective in classification tasks, it may not be the ideal tool to make feature selections when lots of covariates are correlated, especially in modern large p small n problems.

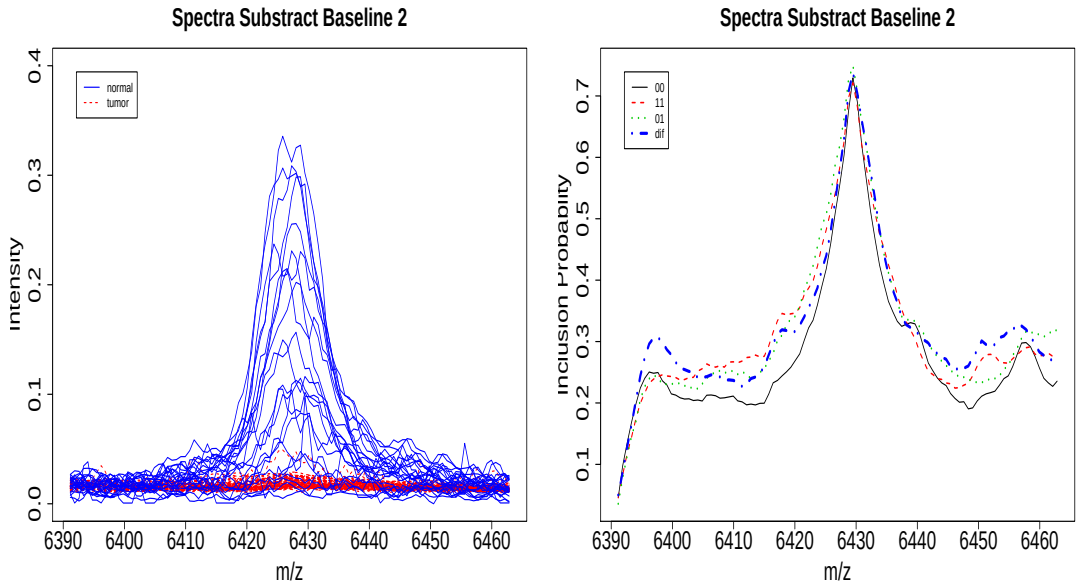

Figure 4.6: Spectra subtracting baseline 2 and the inclusion probabilities.

Section 4.2.3 discussed about the elicitation of the hyper-parameters q_0 and q_1. Alternatively, if we are not certain how to choose a point estimator for q_0 and q_1, we can put a prior distribution for the hyper-parameters. In fact, with independent Beta prior distributions on q_0 and q_1, we have conjugate updates for q_0 and q_1 conditional on other parameters. In practice, we would like to specify the prior distributions for q and s rather than q_0 and q_1, because q and s are easier to interpret. In addition, we prefer specifying independent prior distributions for q and s, because q represents the proportion of 1's and s represents the stickiness of those 1's. This results dependence in prior specification for q_0 and q_1. Because q and s are restricted by $s < (1-q)/q$, we suggest specifying a bivariate logistic normal prior distribution for (q_0, q_1), and then transform it back to (q, s) domain the check whether the prior distributions are appropriate. The bivariate logistic normal distributions are easy to work with, and the correlation for two variables can be induced from the logistic normal covariance matrix. Our preliminary experience with the hyper-prior distribution suggests more

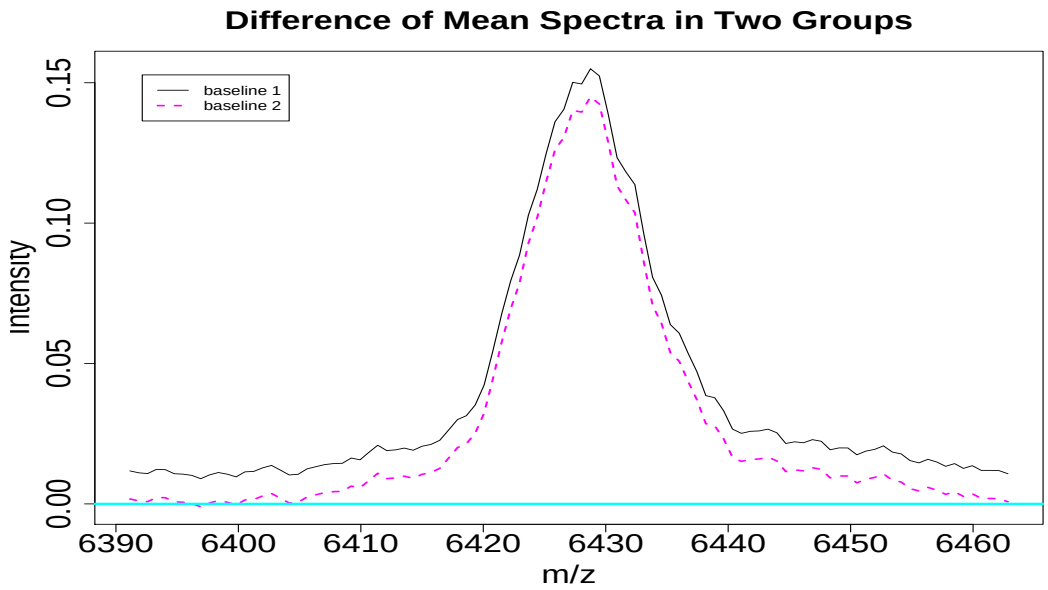

Figure 4.7: The difference of the intensities of the mean spectra in two groups subtracting baseline 1 and 2.

dilution in the inclusion probability for kernel scale parameters.

Another potential research area is to go beyond the Gaussian kernel function. For example, for single nucleotide polymorphism data sets, the value at each location is categorical, *i.e.* A, C, T or G. The kernel function can be constructed with the counting measure, which fits naturally in the BACK framework.

Appendix A

Truncation Approximation for $S\alpha S$ Lévy Random Fields

The symmetric α-stable ($S\alpha S$) Lévy random measure $\nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda})$ is defined on $\mathbb{R} \times \mathbb{X} \times \mathbb{R}_+^r$. Thus it induces a $S\alpha S$ Lévy random field mapping functions $g \in L_2(\mathbb{X} \times \mathbb{R}_+^r)$ to random variables $\mathcal{L}[g] = \int_{\mathbb{X}\times\mathbb{R}_+^r} g(\boldsymbol{\chi}, \boldsymbol{\lambda})\mathcal{L}(d\boldsymbol{\chi}, d\boldsymbol{\lambda})$, with characteristic function

$$\exp\Big\{ \int_{\mathbb{X}\times\mathbb{R}_+^r} g(\boldsymbol{\chi}, \boldsymbol{\lambda}) d\delta_h(\boldsymbol{\chi}, \boldsymbol{\lambda}) + \int_{\mathbb{R}\times\mathbb{X}\times\mathbb{R}_+^r} \left(e^{it\beta g(\boldsymbol{\chi},\boldsymbol{\lambda})} - 1 - it\, h(\beta) g(\boldsymbol{\chi}, \boldsymbol{\lambda})\right) \nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) \Big\} \tag{A.1}$$

where $h(\beta)$ is a compensator function determined uniquely by the signed drift measure $\delta_h(d\boldsymbol{\chi}, d\boldsymbol{\lambda})$ on $\mathbb{X} \times \mathbb{R}_+^r$ (see Tu *et al.*, 2006; Rajput and Rosiński, 1989). For example, when $h_1(\beta) = \sin\beta$, the corresponding drift measure $\delta_{h_1} \equiv 0$. Another convenient compensator function is $h_2(\beta) = \beta 1_{[-1,1]}(\beta)$, and the corresponding drift measure δ_{h_2} satisfies

$$\begin{aligned} \delta_{h_2}(A) &= \delta_{h_1}(A) + \int_{\mathbb{R}\times A} it\,(h_2(\beta) - h_1(\beta))\nu(d\beta, d\omega) \\ &= \int_{\mathbb{R}\times A} it\left(\beta 1_{[-1,1]}(\beta) - \sin\beta\right)\nu(d\beta, d\omega) \end{aligned}$$

for any measurable $A \in \mathbb{X} \times \mathbb{R}_+^r$.

Because $S\alpha S$ Lévy measure $\nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda})$ in (2.6) is not locally β integrable. with compensator function $h(\beta)$, the $S\alpha S$ random variable $mL[g]$ with characteristic

function (A.1). can be constructed by

$$\mathcal{L}[g] = \iint_{\mathbb{R}\times\mathbb{X}\times\mathbb{R}_+^r} [\beta - h(\beta)]\, g(\boldsymbol{\chi},\boldsymbol{\lambda})\mathcal{N}(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) + \iint_{\mathbb{R}\times\Omega} h(\beta)\, g(\boldsymbol{\chi},\boldsymbol{\lambda})\tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) \quad \text{(A.2)}$$

where $mN \sim \mathsf{Po}(\nu)$ is the Poisson random measure on $\mathbb{R}\times\mathbb{X}\times\mathbb{R}_+^r$, and $\tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) = \mathcal{N}(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) - \nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda})$ is the centered Poisson random measure, inducing an isometry from $L_2(\mathbb{R}\times\mathbb{X}\times\mathbb{R}_+^r, \nu)$ to the square integrable zero-mean random variables, (see Sato, 1999, page 38). In particular, take $h(\beta) = h_2(\beta) = \beta 1_{[-1,1]}(\beta)$, (A.2) may be re-written as

$$\mathcal{L}[g] = \int_{[-1,1]^c\times\mathbb{X}\times\mathbb{R}_+^r} \beta g(\boldsymbol{\chi},\boldsymbol{\lambda})\, \mathcal{N}(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) + \int_{[-1,1]\times\mathbb{X}\times\mathbb{R}_+^r} \beta g(\boldsymbol{\chi},\boldsymbol{\lambda})\, d\tilde{\mathcal{N}}(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) \quad \text{(A.3)}$$

Because the SαS Lévy measure in (2.6) is not finite, we cannot sample from the construction (A.3) directly. Instead, approximate the Lévy random field with $\mathcal{L}_\epsilon$, such that

$$\mathcal{L}_\epsilon[g] \;=\; \int_{[-\epsilon,\epsilon]^c\times\mathbb{X}\times\mathbb{R}_+^r} \beta g(\boldsymbol{\chi},\boldsymbol{\lambda})\, d\mathcal{N}. \quad \text{(A.4)}$$

The expected L_2 discrepancy is finite:

$$\begin{aligned} \mathsf{E}\big|\mathcal{L}[g] - \mathcal{L}_\epsilon[g]\big|^2 &= \int_{\mathbb{R}\times\mathbb{X}\times\mathbb{R}_+^r} g(\boldsymbol{\chi},\boldsymbol{\lambda})^2\beta^2 1_{|\beta|\le\epsilon}\nu(d\beta, d\boldsymbol{\chi}, d\boldsymbol{\lambda}) \\ &= ||g||_2^2 \frac{2\gamma\Gamma(\alpha+1)}{\pi(2-\alpha)} \sin\left(\frac{\pi\alpha}{2}\right)\epsilon^{2-\alpha}. \end{aligned}$$

In other words, $\mathcal{L}[g] - \mathcal{L}_\epsilon[g]$ converges to 0 in L_2 as ϵ goes to zero.

Appendix B

Proof of Theorem 1

For any $g \in L_2(\mathbb{X}, \pi(d\boldsymbol{\chi}))$, rewrite the Poisson construction (3.10) as

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

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 (B.1) 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 (B.1) is also zero.

Notice that the difference between (B.1) and (3.11) is

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

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

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

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 (B.2),

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

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

Appendix 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, conditioned on other parameters.

2. Update $\boldsymbol{\lambda}$ using standard Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970), conditioned 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 minimum of 1 and

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

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

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

For classification problems,

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

2. Simulate $\boldsymbol{\beta}^* \sim \mathsf{No}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ is defined in (3.20).

3. Simulate $\mathbf{z}$ from its full conditional distribution given $\mathbf{y}$, $\mathbf{n}$, $\boldsymbol{\beta}^*$, *i.e.*

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

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 $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$, 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 $n_{i^{(prop)}} \neq 0$, 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 minimum of 1 and the product of the conditional posterior density ratio

and the proposal density ratio,

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

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 minimum and

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

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

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

$$\frac{p(y \mid \boldsymbol{\theta}^{(prop)})\mathsf{Ga}(\varphi^{*(prop)}_{i^{(prop)}} \mid \alpha/2, \alpha\epsilon^2/2)\varphi^{*(prop)}_{i^{(prop)}}}{p(y \mid \boldsymbol{\theta})\mathsf{Ga}(\varphi^{*}_{i^{(prop)}} \mid \alpha/2, \alpha\epsilon^2/2)\varphi^{*}_{i^{(prop)}}} \tag{C.1}$$

   (b) **Update $(\mathbf{n}, \boldsymbol{\varphi})$.** 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 $\boldsymbol{\varphi}^*$, otherwise, propose a new φ^* from its prior distribution. The multinominal prior density for $\mathbf{n}$ exactly cancels the proposal density, hence the acceptance rate is the

minimum of 1 and

$$\frac{p(y \mid \boldsymbol{\theta}^{(prop)})}{p(y \mid \boldsymbol{\theta})}. \tag{C.2}$$

Bibliography

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous repsonse data. *J. Am. Stat. Assoc.*, **88**, 669–679.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. John Wiley & Sons.

Besag, J. E. (1972) Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B: Methodological*, **34**, 75–83.

Besag, J. E. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological*, **36**, 192–236.

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.

Cawley, G. C. and Talbot, N. L. C. (2002) Sparse bayesian kernel logistic regression. *Speech Communication*.

Cawley, G. C. and Talbot, N. L. C. (2004) Efficient model selection for kernel logistic regression. *International Conference on Pattern Recognition*, **2**, 439–442.

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.

Chiaromonte, F. and Cook, R. D. (2002) Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, pp. 768–795.

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.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Cleveland, W. S. and Devlin, S. J. (1988) Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, 596–610.

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 rkeerthiepresentations 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. (1998) Principal hessian directions revisited. *J. Am. Stat. Assoc.*, **93**, 84–94.

Cook, R. D. and Lee, H. (1999) Dimension reduction in binary response regression. *J. Am. Stat. Assoc.*, **94**, 1187–1200.

Cook, R. D. and Ni, L. (2005) Sufficient dimension reduction via inverse regression a minimum discrepancy approach. *J. Am. Stat. Assoc.*, **100**, 410–429.

Cook, R. D. and Weisberg, S. (1991) Discussion of a paper by k. c. li. *J. Am. Stat. Assoc.*, **86**, 328–332.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge, UK: Cambridge Univ. Press.

Downs, T. D. (1972) Orientation statistics. *Biometrika*, **59**, 665–676.

Duan, N. and Li, K.-C. (1991) Slicing regression: A link-free regression method. *The Annals of Statistics*, **19**, 505–530.

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. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–374.

Globerson, A. and Tishby, N. (2003) Sufficient dimensionality reduction. *J. Mach. Learn. Res.*, **3**, 1307–1331.

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.

Hastings, W. K. (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.

Higdon, D. M. (1998) Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, **93**, 585–595.

Hoff, P. (2007) Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. Technical Report 517, University of Washington Statistics Department.

Hofmann, T., Schölkopf, B. and Smola, A. J. (2008) Kernel methods in machine learning. *The Annals of Statistics*, **36**, 1171–1220.

Jacod, J. and Shiryaev, A. N. (1987) *Limit Theorems for Stochastic Processes*, vol. 288 of *Grundlehren der mathematischen Wissenschaften*. Berlin, DE: Springer-Verlag.

Johnson, R. (1995) CMU StaLib Datasets Archive. On-line at http://lib.stat.cmu.edu/datasets/.

Khatri, C. G. and Mardia, K. V. (1977) The von Mises-Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society, Series B: Methodological*, **39**, 95–106.

Khinchine, A. Y. and Lévy, P. (1936) Sur les lois stables. *C. R. Acad. Sci. Paris*, **202**, 374–376.

Kingman, J. F. C. (1967) Completely random measures. *Pacific Journal of Mathematics*, **21**, 59–78.

Kwapień, S. and Woyczyński, W. A., eds. (1992) *Random Series and Stochastic Integrals: Single and Multiple*. Probability and its Applications, Boston, MA: Birkhäuser.

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.

Li, F. and Zhang, N. R. (2008) Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Tech. rep., Harvard University. Working Paper.

Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.*, **86**, 316–327.

Li, K.-C. (1992) On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *J. Am. Stat. Assoc.*, **87**, 1025–1039.

Mardia, K. V. (1972) *Statistics of Directional Data*. London and New York: Academic Press.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A. and Kovayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, **21**, 269–292.

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.

Noy, K. and Fasulo, D. (2007) Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics*, **23**, 2528–2535.

Nuñez Antonio, G. and Gutiérrez-Peña, E. (2005) A Bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, **32**, 995–1001.

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.

Presnell, B., Morrison, S. P. and Littell, R. C. (1998) Projected multivariate linear models for directional data. *J. Am. Stat. Assoc.*, **93**, 1068–1077.

Rajput, B. S. and Rosiński, J. (1989) Spectral representations of infinitely divisible processes. *Probab. Theory Rel.*, **82**, 451–487.

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.

Shao, Y., Cook, R. D. and Weisberg, S. (2007) Marginal tests with sliced average variance estimation. *Biometrika*, **94**, 285–296.

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.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A. and Le, Q.-T. (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, **20**, 3034–3044.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **67**, 91–108.

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.

Watson, G. S. (1983) *Statistics on Sphere*. New York, Wiley.

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/∼mlearn/MLRepository.html.

Wolpert, R. L. and Taqqu, M. S. (2005) Fractional Ornstein-Uhlenbeck Lévy processes and the Telecom process: Upstairs and downstairs. *Signal Processing*, **85**, 1523–1545.

Wu, Q., Liang, F. and Mukherjee, S. (2007) Regularized sliced inverse regression for kernel models. Tech. Rep. 07-25, Duke University. Discussion Paper.

Biography

Zhi Ouyang obtained the Bachelor of Science degree in Mathematics at Beijing University in 2004. He has since been at Duke Univeristy in the Ph.D. program of Statistical Science. He obtained the Master of Science degree in Statistics at Duke University in 2007.