BAYESIAN METHODS IN BIOINFORMATICS

A Dissertation

by

VEERABHADRAN BALADANDAYUTHAPANI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2005

Major Subject: Statistics

BAYESIAN METHODS IN BIOINFORMATICS

A Dissertation

by

VEERABHADRAN BALADANDAYUTHAPANI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Commitee,	Raymond J. Carroll
	Bani K. Mallick
Committee Members,	Naisyin Wang
	Rosemary L. Walzem
Head of Department,	Simon J. Sheather

December 2005

Major Subject: Statistics

ABSTRACT

Bayesian Methods in Bioinformatics. (December 2005)

Veerabhadran Baladandayuthapani, B.Sc., Indian Institute of Technology,
Kharagpur India;
M.A., University of Rochester

Co-Chairs of Advisory Committee: Dr. Raymond J. Carroll
Dr. Bani K. Mallick

This work is directed towards developing flexible Bayesian statistical methods in the semi- and nonparamteric regression modeling framework with special focus on analyzing data from biological and genetic experiments. This dissertation attempts to solve two such problems in this area. In the first part, we study penalized regression splines (P-splines), which are low–order basis splines with a penalty to avoid undersmoothing. Such P–splines are typically not spatially adaptive, and hence can have trouble when functions are varying rapidly. We model the penalty parameter inherent in the P–spline method as a heteroscedastic regression function. We develop a full Bayesian hierarchical structure to do this and use Markov Chain Monte Carlo techniques for drawing random samples from the posterior for inference. We show that the approach achieves very competitive performance as compared to other methods. The second part focuses on modeling DNA microarray data. Microarray technology enables us to monitor the expression levels of thousands of genes simultaneously and hence to obtain a better picture of the interactions between the genes. In order to understand the biological structure underlying these gene interactions, we present a hierarchical nonparametric Bayesian model based on Multivariate Adaptive Regression Splines (MARS) to capture the functional relationship between genes and also between genes and disease status. The novelty of the approach lies in the attempt to capture the complex nonlinear dependencies between the genes which could otherwise be missed by linear approaches. The Bayesian model is flexible enough to identify significant genes of interest as well as model the functional relationships between the genes. The effectiveness of the proposed methodology is illustrated on leukemia and breast cancer datasets.

To Mom, Dad, Ramesh and Upali

ACKNOWLEDGEMENTS

I am indebted to many people for helping me attain my Ph.D. I attribute my success to the positive interactions I have had with my family, friends, and various mentors. I feel very fortunate that I have been able to surround myself with so many people who have had a significant impact on my life.

First, I am greatly indebted to my advisors Drs. Raymond Carroll and Bani Mallick. I am honored to have had the opportunity to be Dr. Carroll's student; it has been a privilege to get a glimpse of the field of statistics through his eyes. Dr. Carroll has taught me to be a self-sufficient researcher; I will always be grateful to him. I am privileged to have worked with Dr. Mallick who with his unfailing energy and drive, has been a professional mentor above and beyond the role of an academic advisor.

Additionally, I would like to thank my parents and brother. Their emotional support and encouragement over the years has been immeasurable. Their willingness to make sacrifices on my behalf has always inspired me to try to achieve more than I thought was possible.

I would also like to thank the members of my committee: Dr. Naisyin Wang and Dr. Rosemary Walzem. I appreciate all of the time they took out of their busy schedules to attend my preliminary examination and dissertation defense. I would like to thank Dr. Wang for always making time to answer my questions despite her busy schedule.

Lastly and most importantly, I want to express my warmest thanks to my wife Upali. Without her friendship, emotional support, and unflinching help none of this may have been possible. I am unbelievably fortunate to have someone so inspirational in my life. I am extremely thankful to her for making my graduate school experience infinitely more pleasant. It would be an understatement that this work would not have seen the light of the day but for her.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

This dissertation introduces Bayesian nonparametric regression modeling tools and utilizes modern Markov Chain Monte Carlo (MCMC) techniques to explore the posterior distributions of interest induced by the models. Bayesian methodology has generated immense interest due to two basic reasons. First, Bayesian methods take an axiomatic view of uncertainity allowing the user to make coherent inference and second, Bayesian modeling is particularly well suited to incorporating prior information, which is often available. Special focus is on developing Bayesian statistical machinery for modeling data from functional genomics.

In Chapter II, we study penalized regression splines (P-splines), which are low–order basis splines with a penalty to avoid undersmoothing. Such P–splines are typically not spatially adaptive, and hence can have trouble when functions are varying rapidly. Our approach is to model the penalty parameter inherent in the P–spline method as a heteroscedastic regression function. We develop a full Bayesian hierarchical structure to do this and use Markov Chain Monte Carlo techniques for drawing random samples from the posterior for inference. The advantage of using a Bayesian approach to P–splines is that it allows for simultaneous estimation of the smooth functions and the underlying penalty curve in addition to providing uncertainty intervals of the estimated curve. The Bayesian credible intervals obtained for the estimated curve are shown to have pointwise coverage probabilities close to

The format and style follow that of *Journal of the American Statistical Association*.

nominal. The method is extended to additive models with simultaneous spline based penalty functions for the unknown functions. In simulations, the approach achieves very competitive performance with the current best frequentist P–spline method in terms of frequentist mean squared error and coverage probabilities of the credible intervals, and performs better than some of the other Bayesian methods.

Chapter III deals with DNA microarray data. DNA microarray technology enables us to monitor the expression levels of thousands of genes simultaneously, and hence to obtain a better picture of the interactions between the genes. In order to understand the biological structure underlying these gene interactions, we present here a statistical approach to model the functional relationship between genes and also between genes and disease status. We suggest a hierarchical Bayesian model based on Multivariate Adaptive Regression Splines (MARS) to model these complex nonlinear interaction functions. The novelty of the approach lies in the fact that we attempt to capture the complex nonlinear dependencies between the genes which otherwise would have been missed by linear approaches. Owing to the large number of genes (variables) and the complexity of the data, we use MCMC based stochastic search algorithms to choose among models. The Bayesian model is flexible enough to identify significant genes as well as model the functional relationships between them. The effectiveness of the proposed methodology is illustrated using two publicly available microarray data sets: Leukemia and hereditary breast cancer.

Chapter IV provides a summary of the results in this dissertation and some open questions are posed for future research.

CHAPTER II

SPATIALLY ADAPTIVE BAYESIAN PENALIZED REGRESSION SPLINES[*]

2.1 Introduction

Regression splines are approximations to functions typically using a low–order number of basis functions. Such splines, like all splines, are subject to a lack of smoothness and various strategies have been proposed to attain this smoothness. A particularly appealing class are the regression P–splines (Eilers and Marx 1996), which achieve smoothness by penalizing the sum of squares or likelihood by a single penalty parameter. The penalty parameter and the fit using P–splines are easy to compute using mixed model technology (see Robinson 1991; Coull, Ruppert and Wand 2001; Rice and Wu 2001, among others), and are not sensitive to knot parameter selection (Ruppert 2002).

Despite these advantages, P–splines with a single penalty parameter are not suitable for spatially adaptive functions that can oscillate rapidly in some regions and are rather smooth in other regions (Wand 2000). Rather than using a global penalty parameter, Ruppert and Carroll (2000) proposed a local penalty method wherein the penalty is allowed to vary spatially so as to adapt to the spatial heterogeneity in the regression function. The web site http://orie.cornell.edu/∼davidr contains the MATLAB code for computing this spatially adaptive estimator.

The purpose of this chapter is to construct a Bayesian version of the local penalty method. We do this by modeling the penalty as another regression P–spline, in effect a variance function, in a hierarchical structure. The method is rel-

[*] Reprinted with permission from the *Journal of Computational and Graphical Statistics*. Copyright 2005 by the American Statistical Association. All rights reserved.

atively simple to compute and implement, and the MATLAB code for it is given at http://stat.tamu.edu/∼veera. The advantage of using a Bayesian approach to P–splines is that it allows for simultaneous estimation of the function and the underlying penalty curve in addition to providing uncertainty intervals for the estimated curve. We show that our method achieves competitive performance with that of Ruppert and Carroll in terms of frequentist mean squared error and coverage probabilities of credible intervals. The Bayesian credible intervals obtained for the estimated curve are shown to have pointwise frequentist coverage probabilities close to nominal. In simulations our method outperforms, sometimes substantially, many other Bayesian methods existing in literature.

The chapter is structured as follows: Section 2.2 introduces the Bayesian model used, along with the prior and distributional assumptions on the random variables and parameters. Section 2.3 is devoted to the MCMC setup for the calculations. Section 2.4 discusses the simulation study undertaken and the results of our findings. We extend the univariate ideas to additive models in Section 2.5. Technical details are collected into Appendix A

2.2 Model Formulation

Given data (X_i, Y_i), where X_i is univariate, our nonparametric model is defined by

$$Y_i = m(X_i) + \epsilon_i,$$

where $m(\bullet)$ is an unknown function, the ϵ_i's are independent conditional on X_i and normally distributed with mean zero and variance σ_Y^2.

To estimate $m(\bullet)$ we use regression P–splines. As the basis functions, here we use piecewise polynomial functions whose highest order derivative takes jumps at fixed "knots". Other basis functions such as B-splines (de Boor 1978) could also be used.

With this basis, the functional form of the regression spline of degree $p \geq 1$ is given by

$$m(X) = \alpha_{Y0} + \alpha_{Y1}X + \ldots + \alpha_{Yp}X^p + \sum_{j=1}^{M_Y} \beta_{Yj}(X - \kappa_{Yj})_+^p,$$

where $(\alpha_{Y0}, \ldots, \alpha_{Yp}, \beta_{Y1}, \ldots, \beta_{YM_Y})$ is a vector of regression coefficients and $(a)_+^p = a^p I(a \geq 0)$, and $\kappa_{Y1} < \ldots < \kappa_{YM_Y}$ are fixed knots.

To model the unknown smooth function $m(\bullet)$, we illustrate the theory using regression splines of degree 1, so that

$$m(X) = \alpha_{Y0} + \alpha_{Y1}X + \sum_{j=1}^{M_Y} \beta_{Yj}(X - \kappa_{Yj})_+, \tag{2.1}$$

Of course, changes to polynomials of higher degree are trivial. We take M_Y, the number of knots, to be large but much less than n, the number of data points. Unlike knot-selection techniques we retain all candidate knots. In this particular method, we take the knots to be the equally spaced sample quantiles of X, although one could just as easily take the knots to be equally spaced.

The number of knots here is specified by the user. Although the choice is not crucial (Ruppert 2002) a minimum number knots are needed to capture the spatial variability in the data. The choice of knots is discussed in detail later in the chapter (Section 2.4.2).

We can interpret (2.1) as a Bayesian linear model. Rewrite (2.1) as

$$Y = Z_Y \Omega_Y + \epsilon_Y, \tag{2.2}$$

where $Y_{n\times1} = (Y_1, \ldots, Y_n)^T$, $\Omega_Y = (\alpha_{Y0}, \alpha_{Y1}, \beta_{Y1}, \ldots, \beta_{YM_Y})^T$ is a $(M_Y + 2) \times 1$ vector of regression coefficients, $\epsilon_Y = (\epsilon_1, \ldots, \epsilon_n)^T$ is $n \times 1$ error vector and the design matrix $\mathbf{Z_Y}$ is defined as

$$\mathbf{Z_Y} = \begin{bmatrix} 1 & X_1 & (X_1 - \kappa_{Y1})_+ & \dots & (X_1 - \kappa_{YM_Y})_+ \\ 1 & X_2 & (X_2 - \kappa_{Y1})_+ & \dots & (X_2 - \kappa_{YM_Y})_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & (X_n - \kappa_{Y1})_+ & \dots & (X_n - \kappa_{YM_Y})_+ \end{bmatrix}.$$

Suppose that $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed Normal$(0, \sigma_Y^2)$. The parameters in $(\alpha_{Y0}, \alpha_{Y1})$ can be considered as fixed effects in the model. We put a normal prior on $(\alpha_{Y0}, \alpha_{Y1})$ with 0 mean and large variance (say 100). This effectively acts as a non-informative uniform prior on the fixed effects. The random variables in $\{\beta_{Yj}\}_{j=1}^{M_Y}$, are assumed a priori independent and normally distributed, i.e., $\beta_{Yj} \sim$ Normal$\{0, \sigma_j^2(\kappa_{Yj})\}$, where $j = 1, \dots, M_Y$. Note here that $\sigma_j^2(\kappa_{Yj})$ is the smoothing parameter (shrinkage or ridge parameter).

In the usual regression P–spline formulation with a global smoothing parameter, the $\sigma_j^2(\kappa_{Yj})$ are all constant as a function of j, so that the smoothing is not spatially adaptive. We next describe how we extend the model to allow for spatially adaptive smoothing.

To develop a spatially adaptive technique we need to model $\sigma_j^2(\kappa_{Yj})$. This is crucial to capturing the spatial heterogeneity of the data because different smoothing parameters lend different amounts of smoothing in different regions. Allowing the smoothing parameter to be spatially adaptive also helps improve the mean squared error (MSE) of the fits, as well as the accuracy of inference (Ruppert and Carroll 2000; Wand 2000). In this spirit, we develop a hierarchical model for $\sigma_j^2(\kappa_{Yj})$, where $\sigma^2(\bullet)$ is a function evaluated at the knots (κ_{Yj}). The functional form of $\sigma^2(\bullet)$ is taken to be another linear regression spline, e.g., for a linear spline

$$-\log\{\sigma^2(X)\} \quad = \quad \alpha_{s_0} + \alpha_{s_1} X + \sum_{k=1}^{M_s} \beta_{s_k} (X - \kappa_{s_k})_+, \qquad (2.3)$$

where again $\kappa_1 < \ldots < \kappa_{M_s}$ are fixed knots. The number of sub-knots M_s is again user specified and is typically far less than M_Y, the number of knots in the original spline. The knots $\{\kappa_k\}_{k=1}^{M_s}$ are again taken to be equally spaced quantiles of X. We now write (2.3) as a Bayesian linear model:

$$\rho = Z_s \Omega_s \tag{2.4}$$

where $\rho = [-\log\{\sigma^2(\kappa_1)\}, \ldots, -\log\{\sigma^2(\kappa_{M_Y})\}]^T$, $\Omega_s = (\alpha_{s_0}, \alpha_{s_1}, \beta_{s_1}, \ldots, \beta_{s_{M_s}})^T$ is an $(M_s + 2) \times 1$ vector and Z_s is the design matrix, identical to that for (2.2) except the change in the knots.

The random variables in the above equation are again assumed a priori independent and normally distributed, i.e., $\beta_{s_k} \sim \text{Normal}(0, \xi^2)$, where $k = 1, \ldots, M_s$ and the parameters $(\alpha_{s_0}, \alpha_{s_1})$ are again independent and normally distributed with zero mean and large variance.

As described in Section 2.3, although the motivation as a variance function to achieve spatially adaptive smoothing is clear, we will actually use a slight modification of (2.3)–(2.4) in order to avoid Ω_s having to be sampled by a complex Metropolis–Hastings step.

2.3 Implementation via Markov Chain Monte Carlo Simulation

In this section we set up the framework to carry out the Markov Chain Monte Carlo (MCMC) calculations. The prior distributions of the variance (σ_Y^2) of the error vector ϵ_Y, and ξ^2, the variance of the β_{s_k}'s, are taken to be a conjugate inverse gamma distribution with parameters (a_Y, b_Y) and (a_s, b_s) respectively, i.e., $\sigma_Y^2 \sim IG(a_Y, b_Y)$ and $\xi^2 \sim IG(a_s, b_s)$, where IG($\bullet$) is the inverse gamma distribution.

The parameters and random variables to be estimated in the model are $\Omega_Y, \Omega_s, \xi^2$ and σ_Y^2. With the above model and prior set-up all the conditional distributions turn out to be of known standard forms except that of Ω_s, which is a complex multivariate density. Hence we need a multivariate Metropolis-Hastings (MH) step to generate the samples. Since this involves searching over a $(M_s + 2)$-dimensional space for convergence, we noticed during the simulations that the movement of the MH step was very slow.

Hence we resort to the following device to reduce the dimension, thereby making the moves faster. We add an error term (ϵ_u) to the functional form of $\sigma^2(X)$ in (2.3)–(2.4), leading to the model

$$\rho = Z_s\Omega_s + \epsilon_u, \tag{2.5}$$

where $\epsilon_u = \text{Normal}(0, \sigma_u^2 I)$. We fix the value of σ_u^2 for our simulations to $= 0.01$ because this variance is unidentified in the model. This device reduces the computational costs by reducing the MH step to one dimension to generate each of $\sigma_j^2(\kappa_{Yj})$'s, which are now conditionally dependent only on Ω_s and conditionally independent of the rest of the parameters. This in effect makes the movement of the MCMC samples across the model space extremely fast and also improves the acceptance rate of MH moves. In our simulations we found that the choice of the value of σ_u^2 does not have great influence on the performance of the MCMC. The complete conditional posteriors are derived in the Appendix A.

2.4 Simulations

In this section we present simulation studies primarily to evaluate the frequentist performance of our methodology and to compare it with other related approaches in literature. Section 2.4.1 compares the Bayesian P–spline approach to the frequentist local penalty approach of Ruppert and Carroll (2000) and with a variety of recent Bayesian approaches, in particular with the BARS (Bayesian Adaptive Regression

Splines) method proposed by DiMatteo, Genovese and Kass (2001). Section 2.4.2 discusses the issue of the choice of knots in the implementation of our algorithm.

2.4.1 Comparison with Other Methods

We compare our Bayesian approach with the frequentist penalized splines approach (RC, Ruppert and Carroll 2000), through the following simulation study. The X's were equally spaced on $[0,1]$, $n = 400$, $\sigma_u^2 = 0.01$ and the ϵ_i's were Normal$(0, 0.04)$. First, we use the regression function as in RC whose spatial variability was controlled by parameter j,

$$m(x) = \sqrt{x(1-x)} \sin\left[\frac{2\pi(1 + 2^{(9-4j)/5})}{x + 2^{(9-4j)/5}}\right], \tag{2.6}$$

where $j = 3$ gives low spatial variability and $j = 6$ gives severe spatial variability; see Figure 1 panels (a) and (b). The fits obtained by our algorithm using a truncated power basis function of degree 2 are shown in panels (c) and (d) along with associated 95% credible intervals. The credible intervals are estimated by computing the respective quantiles of the sampled function evaluations.

In this chapter, we allow the smoothing/penalty parameter to be a function of the independent variable $\mathbf{X}$ as in (2.3). As mentioned before, this is important in capturing the spatial heterogeneity in the data by allowing different amounts of smoothing in different regions. We plot the underlying penalty function, $\sigma^2(X)$ in Figure 1 panels (e) and (f). We would expect the value of $\sigma^2(X)$ to be large if the regression curve has rapid changes in curvature, so that the second derivative of the fitted spline can take jumps large enough to accommodate these changes. Conversely, if the curvature changes slowly, then we would expect $\sigma^2(X)$ to be small. Observe that the penalty curve adapts to the spatial heterogeneity of the underlying regression function with large values in the regions where the curve is non-smooth and small

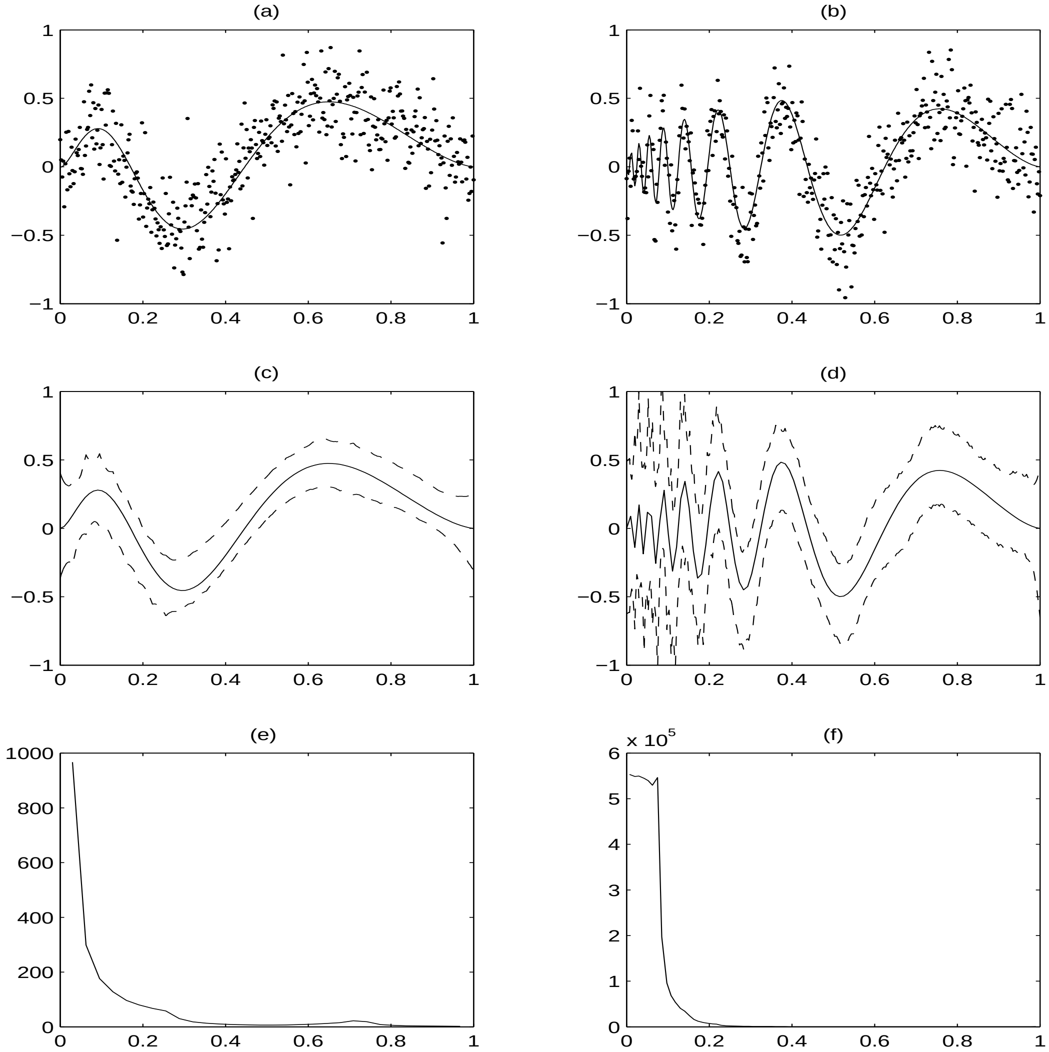


Figure 1. (a) Plot of the test curve used for simulations. The spatial variability of this curve is controlled by the parameter j. In this case j=3 gives low spatial variability. (b) j=6 gives severe spatial variability. (c) The true function with error added. (d) Same as (c) but j = 6. (e) Plot of the estimated regression function. The number knots (M_Y) is 30 and number of subknots (M_s) is 5.(f) Same as (e) but j = 6. Here M_Y=90 and M_s=15. Also shown on the plots are the 95% credible intervals on the fitted curve.

values in smooth regions.

In order to compare the performance of fit we compute the averaged mean squared error (AMSE), which is given by

$$\text{AMSE} = n^{-1}\sum_{i=1}^{n}\left\{\widehat{m}(x_i) - m(x_i)\right\}^2 . \tag{2.7}$$

Our estimated AMSE for $j = 3$ and $j = 6$ is 0.0.0006 and 0.0027 respectively, which is comparable to the those obtained by RC on the same data set, which were 0.0007 and 0.0026 respectively.

We also compared the frequentist coverage properties of the Bayesian credible intervals with the frequentist local penalty confidence intervals of RC and with BARS. BARS employs free-knot splines, where the number and location of knots are random, and uses reversible jump MCMC (Green 1995) for implementation. We consider a spatially heterogeneous regression function,

$$m(X) \quad = \quad \exp\{-400(X-0.6)^2\} + \frac{5}{3}\exp\{-500(X-0.75)^2\} + 2\exp\{-500(X-0.9)^2\}, \tag{2.8}$$

The X's are equally spaced on $[0,1]$, the sample size was $n = 1000$, and the ϵ_i were normally distributed with $\sigma = 0.5$. We use truncated power basis function of degree 2 with $M_Y = 40$ knots and $M_s = 4$. We again set $\sigma_u = 0.01$. The BARS program was graciously provided by the authors of DiMatteo et al. (2001). The BARS estimates are based on a Poisson prior with mean 6 for the number of knots, and the MCMC chain was run for 10,000 iterations with a burn-in period of 1000.

Figure 2 shows a typical data set with the true and fitted function plotted. In order to compare the coverage probabilities of the Bayesian credible intervals, we compute the frequentist coverage probabilities of the 95% credible intervals over

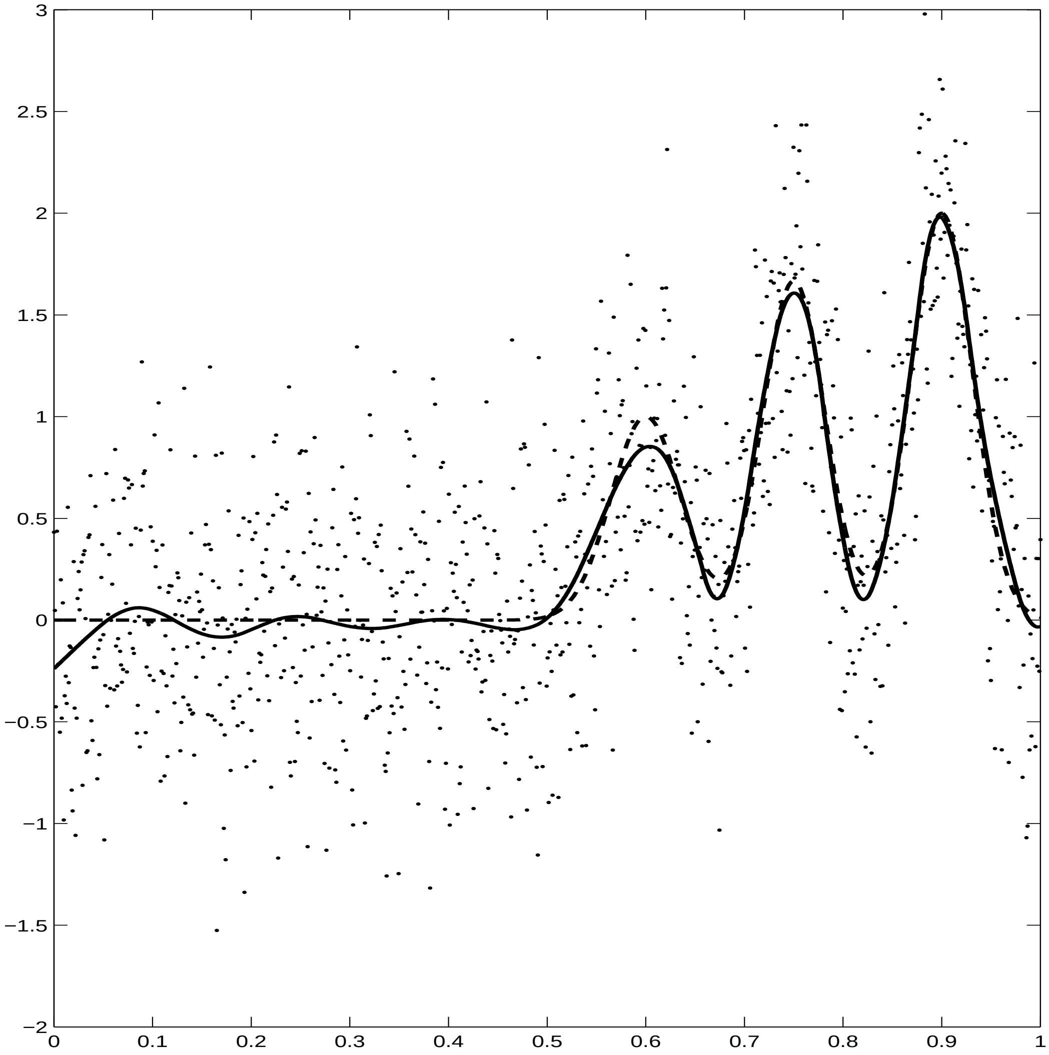


Figure 2. A copy of simulated data for comparing coverage probabilities. Shown are dots = data, dashed curve = true function and solid curve = Bayesian p-spline estimate.

500 simulated data sets. Figure 3 shows the pointwise coverage probabilities of the 95% Bayesian credible intervals along with the "adjusted" local penalty confidence intervals of RC and those obtained by BARS. The adjustment used by RC is to multiply the pointwise posterior variances of the local-penalty estimates by a constant so that the average pointwise posterior variance of the estimate is the same for the global and local penalty estimate (Ruppert and Carroll 2000, Section 4). The coverage probabilities shown have been smoothed using P–splines to remove the Monte Carlo variability. The average coverage probability obtained by the three methods (Bayesian P–splines, RC, BARS) are (95.22%, 96.28%, 94.72%) respectively. The coverage probabilities for both Bayesian P–splines and BARS are slightly closer to the nominal coverage of 95% than the more conservative local penalty intervals of RC. Figure 4 shows the pointwise AMSE using the three methods. The average MSE for BARS (0.0043) is somewhat smaller than the MSE for the Bayesian P-spline (0.0061) and RC (0.0065). Thus, our results are competitive to BARS in terms of frequentist coverage probabilities but BARS does seem to do a slightly better job than our method in terms of overall MSE.

We also compared our method with two other Bayesian approaches, "Automatic Bayesian Curve Fitting" method proposed by Denison, Mallick and Smith (1998a) and the wavelet methods of Donoho and Johnstone (1994): further discussion of some potential problems with the method of Denison, Mallick and Smith is given by DiMatteo et al. (2001). We used the four test curves: 'Heavisine', 'Blocks', 'Bumps' and 'Doppler' as in Donoho and Johnstone . The X's were again equally spaced on $[0,1]$, n was 2048, and the ϵ_i's were Normal$(0,1)$. The values of (M_Y, M_s) are $\{(60,10),(300,30),(90,15),(250,80)\}$ for Heavisine, Blocks, Bumps and Doppler respectively. Denison et al. (1998a) reported the average MSE from 10 replications using the above examples and compared the results with those obtained by Donoho

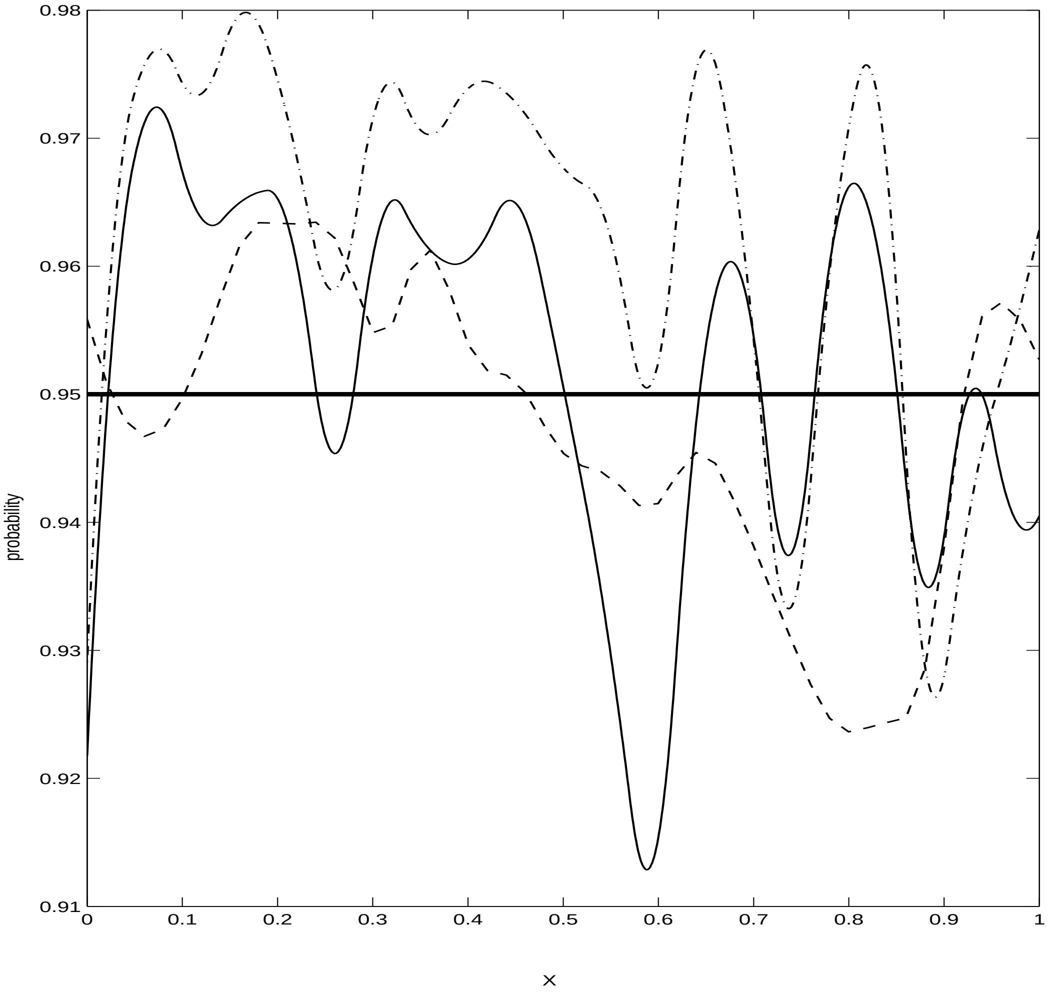


Figure 3. Comparison of coverage probabilities. (a) Shows the frequentist coverage probabilities of 95% credible intervals. The dashed line is using the adjusted local penalty confidence interval of Ruppert and Carroll (2000) and solid line is using the Bayesian credible intervals. The coverage probabilities shown have been smoothed using p-splines to remove the Monte Carlo variability (b) Same as (a) for 90% coverage.

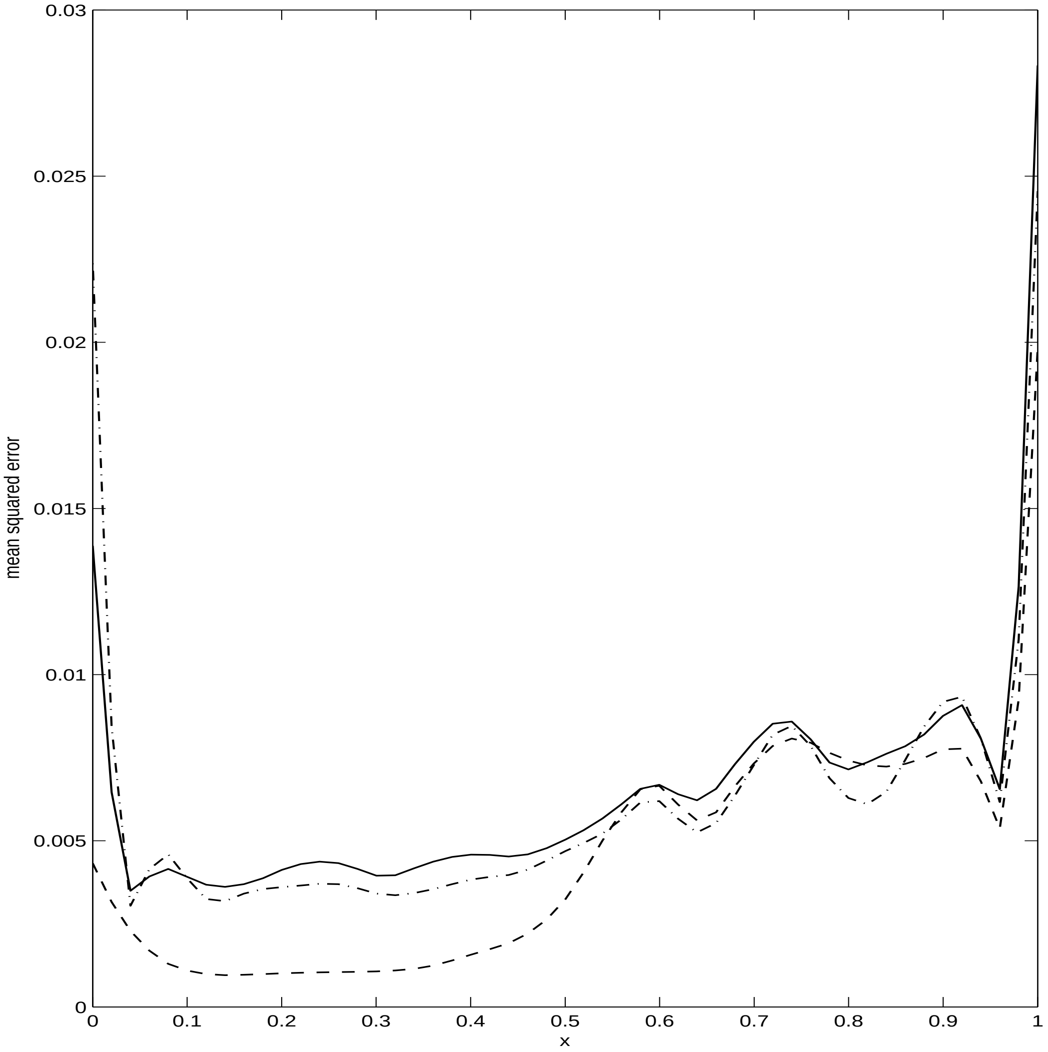


Figure 4. Comparison of pointwise mean squared errors (MSE). The dashed line is using BARS method of (Dimatteo et al. 2001) and solid line is using the Bayesian P-spline method. The coverage probabilities shown have been smoothed using P-splines to remove the Monte Carlo variability.

and Johnstone. Table 1 compares our results to those obtained by Denison et al. (1998a) and Donoho and Johnstone . Here λ_n^* is the optimal wavelet threshold chosen specifically for each data set, while $\{2\log(n)\}^{1/2}$ is a universal threshold proposed by Donoho and Johnstone . As noted in Denison et al. (1998a) the wavelet results are obtained with σ^2 known and, for ease of computation, require the number of data points to be a power of 2. Specifically, we take $n = 2048$ and $\sigma_u^2 = 0.01$ to compare out results with that of Denison et al. (1998a) and Donoho and Johnstone. Our method performs markedly better than the wavelet threshold methods in all the examples considered. Our results are comparable with those obtained by Denison et al. (1998a), for the 'Heavisine', 'Blocks' and 'Bumps' functions but is much better for the 'Doppler' example.

Table 1. Average mean squared error(AMSE) comparison from 10 replications for different example curves across different methods: wavelet threshold methods, Automatic Bayesian curve fitting and Bayesian P-splines

Function	Wavelet threshold λ_n^*	Wavelet threshold $\{2\log(n)\}^{1/2}$	Automatic curve fitting	Bayesian P-splines
Heavisine	0.060	0.083	0.033	0.028
Blocks	0.427	0.905	0.170	0.137
Bumps	0.499	1.080	0.167	0.098
Doppler	0.151	0.318	0.135	0.024

2.4.2 Choice of Knots

In this chapter we present a penalty approach which is similar in spirit to smoothing splines, but with fewer knots. In P-splines the crucial parameter in controlling the amount of smoothness is the penalty, i.e., in our case $\sigma^2(\kappa)$. Once a certain minimum number of knots is reached, further increase in the number of knots causes little change to the fit given by P–spline (Ruppert 2002 ; Ruppert, Wand and Carroll 2003). To this effect we ran an analysis with different number of knots, but the same selection for each method. The X's were equally spaced on $[0,1]$, $n = 400$, $\sigma_u^2 = 0.01$ and the ϵ_i's were Normal$(0, 0.04)$. We again use the regression function as in (2.6) with $j = 3$ (low spatial variability) and $j = 6$ (severe spatial variability). We used 5 different sets of knots for the regression curve and the penalty curve i.e. $\{(20,3),(40,4),(60,6),(90,9),(120,15)\}$. To compare the performance of fit across the different sets of knots we compute the AMSE as in (2.7).

Table 2 shows the AMSE for the test cases described above. For $j = 3$ there is essentially no improvement on the fit of the curve on increasing the number knots. For the severe spatially variable case ($j = 6$) the AMSE improves appreciably by increasing the number of knots from (20,3) to (40,4) but marginally by increasing the knots further. In all the examples we consider, there is evidence that there is a minimum necessary number of knots to be reached to fit the features in the data, and a further increase in the number of knots does not have appreciable effect on the fit. Thus, if enough knots are specified, adaptive P–splines will be able to track the sudden changes in the underlying function, and where the underlying function is smooth the penalty will shrink the jumps at those knots to 0.

For the penalty curve $\sigma^2(X)$, the number of subknots M_s is taken to be much smaller than M_Y, the number of knots for the original regression spline. We tried a variety of choices for M_s in our simulation examples and found that the choice of M_s has relatively little effect on the fit. We keep the value of M_s large enough for the penalty curve to be spatially variable and small enough to reduce the computational cost. In all our simulation examples we take the value of M_s to be less than a sixth of the number of knots chosen for original regression spline (M_Y).

The variance of the error term in all the simulation examples was taken so that we could mimic the simulation setup of the methods to which we compare our method to. In order to study the performance of our estimator in the presence of increased noise we ran a further simulation study. We took the same simulation curve as in (2.1) with $j = 3$, the X's were equally spaced on $[0, 1]$, $n = 400$ and $\sigma_u^2 = 0.01$. The variance of the error term (σ_Y^2) was taken to be at three different levels: (0.04, 0.1, 0.5). The average MSE over 25 simulated datasets was found to be (0.0015, 0.0055, 0.0070) respectively showing that the fitted curve can estimate the underlying regression function well even under increased noise.

Table 2. Average mean squared error(AMSE) comparison using different sets of knots (M_Y, M_s). Shown are the AMSE obtained for two test cases of a simulation example curve (2.6) (see text) where j=3 gives low spatial variability and j=6 gives severe spatial variability

Knot set	$j = 3$	$j = 6$
(20,3)	0.0007	0.0094
(40,4)	0.0007	0.0048
(60,6)	0.0008	0.0036
(90,9)	0.0009	0.0028
(120,15)	0.0012	0.0027

2.5 Extension to Additive Models

2.5.1 An Algorithm for Additive Models

To this point, we have confined our attention to univariate cases only. The methodology developed previously can be easily extended to additive models. The general additive model problem is to find functions m_j such that

$$Y = \alpha + \sum_{j=1}^{p} m_j(X_j) + \epsilon, \tag{2.9}$$

where the X_j are the predictor variables, $E(\epsilon|X_1, ..., X_p) = 0$ and $\text{var}(\epsilon|X_1, ..., X_p) = \sigma_Y^2$. Thus the overall regression function is a sum of p univariate functions, or curve fits. The univariate functions can be modelled with univariate splines, as we shall assume here. A model of the general type in (2.9) is known as an additive model. Hastie and Tibshirani (1990) provide an extensive account of these models.

The extension to Bayesian additive models is straightforward when we use a basis function representation (such as P-splines) for the individual curves. That is, we can write $g(X) = \sum_{j=1}^{p} m_j(X_j)$ in (2.9) as a linear combination of P-splines and regression coefficients as:

$$g(X) = \alpha_0 + \sum_{j=1}^{p} \alpha_{1j} X_j + \sum_{j=1}^{p} \sum_{i=1}^{M_j} \beta_{ji}(X_j - \kappa_{ji})_+, \tag{2.10}$$

where again X_j is the jth predictor in X and M_j is the number of knots for the jth curve. Each one-dimensional function is again described by the parameters β_{ji} (the coefficients) and κ_{ji} (the knots).

As in previous sections we can use the same Bayesian linear model results, to make posterior inference for additive models. Thus the fact that a general set of predictors is now a vector, rather than just a scalar, is of little consequence. In matrix notation we again write

$$Y = B\beta + \epsilon,$$

with $\epsilon \sim \text{Normal}(0, \sigma^2 I)$, $\beta = (\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_{\mathbf{1}}, \ldots, \beta_{\mathbf{p}})^T$ with $\beta_{\mathbf{j}} = (\beta_{j,1}, \ldots, \beta_{j,M_j})$ and

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} 1 & X_1 & B_{1,1}(X_1) & \ldots & B_{1,M_1}(X_1) & B_{2,1}(X_1) & \ldots & B_{p,M_p}(X_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & B_{1,1}(X_n) & \ldots & B_{1,M_n}(X_n) & B_{2,1}(X_n) & \ldots & B_{p,M_p}(X_n) \end{bmatrix}, \\ &= [1 \; X \; \mathbf{B_1} \; \cdots \; \mathbf{B_p}] \end{aligned}$$

where

$$\mathbf{B_j} = \begin{pmatrix} (X_{j1} - \kappa_{j,1})_+ & \dots & (X_{j1} - \kappa_{j,M_j})_+ \\ \vdots & \ddots & \vdots \\ (X_{n1} - \kappa_{j,1})_+ & \dots & (X_{n1} - \kappa_{j,M_j})_+ \end{pmatrix}.$$

With the above formulation, we adopt the same methodology as discussed in the previous sections for the univariate case. The distributional assumptions and prior structure on the random variables and parameters respectively are exactly the same as described in Section 2.2. The functional form of the variance of β is again a linear regression spline as in (2.3).

2.5.2 Simulations of an Additive Model

We take a slightly modified example from Hastie and Tibshirani (1990, pp. 247–251). We simulated from the functions m_1 and m_2 for the model,

$$Y_i = m_1(X_i) + m_2(Z_i) + \epsilon_i, \quad i = 1, \dots, 100,$$

where

$$m_1(X) = \begin{cases} -2X & \text{for } X < 0.6, \\ -1.2 & \text{otherwise,} \end{cases}$$

$$m_2(Z) = \frac{\cos(5\pi Z)}{1 + 3Z^2},$$

with X_i and Z_i generated independently from the Uniform $(0, 1)$ distribution and ϵ_i from an Normal$(0, 0.25)$ distribution. Figure 5 shows the estimates for functions $m_1(X)$ and $m_2(Z)$, along with the credible intervals. The fits are better in terms of AMSE than the estimates provided by Denison et al. (1998a). Also Denison et al. (1998a) used plug-in estimates for the regression coefficients (β's), and thus underestimate the uncertainty. We perform a full Bayesian analysis where in we draw the regression coefficients from the sampler, and hence obtain standard Bayesian credible intervals.

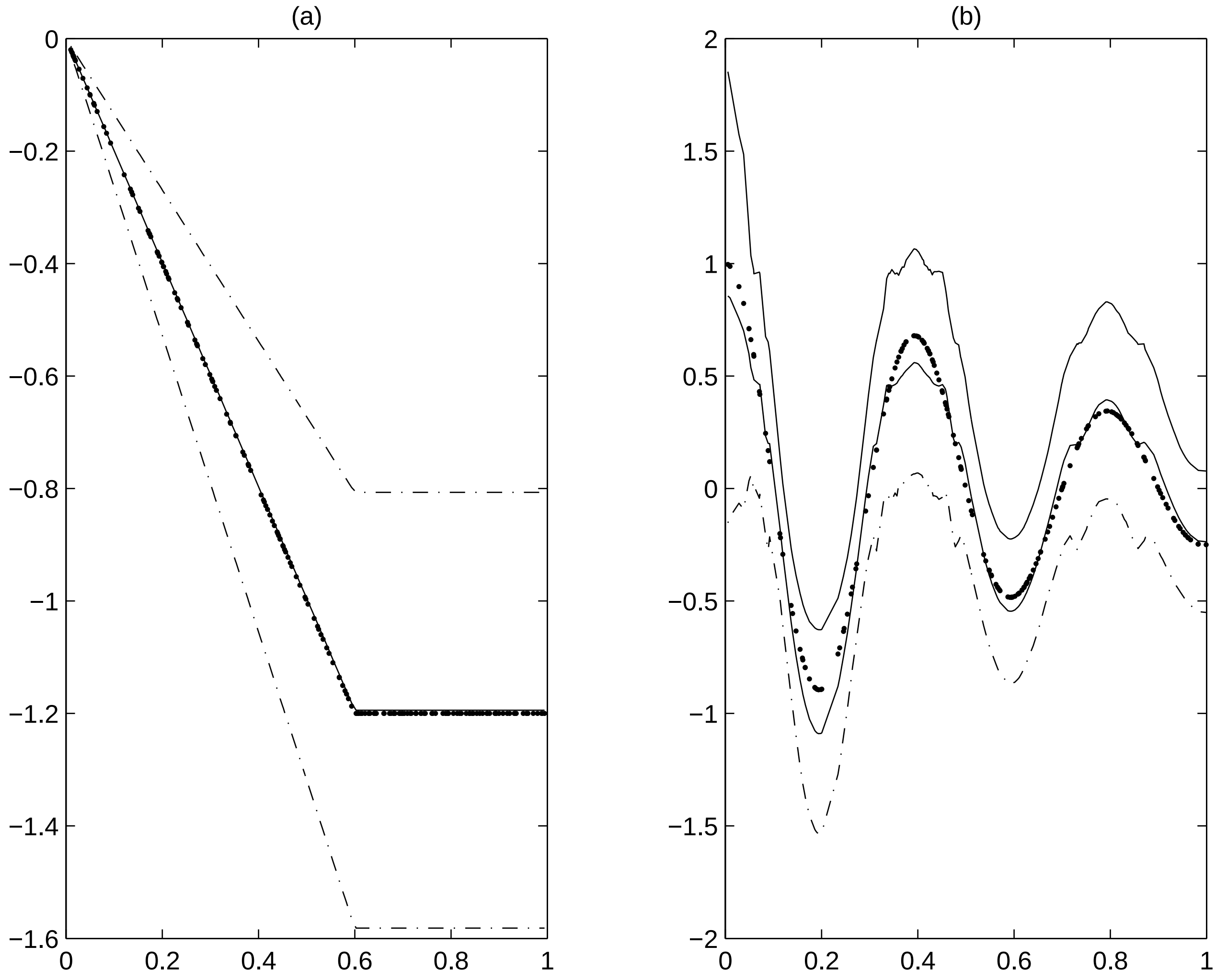


Figure 5. Additive model example. The dotted line represents the true function and the solid line represents the estimate regression function. Also shown are the 95% credible intervals. (a) $m_1(X)$*;(b)* $m_2(Z)$*.*

CHAPTER III

MODELING NONLINEAR GENE INTERACTIONS USING BAYESIAN MARS

3.1 Introduction

DNA microarray technology has revolutionized biological and medical research. The use of DNA microarrays allows simultaneous monitoring of the expressions of thousands of genes (Duggan et al. 1999; Schena et al. 1995), and has emerged as a tool for disease diagnosis. This technology promises to monitor the whole genome on a single chip so that researchers can have a better picture of the interactions among thousands of genes simultaneously. In order to understand the biological structure underlying the gene interactions, i.e., on what scale can we expect genes to interact with each other, we need to model the functional structure between the genes. However, due to the complexity of the data and the curse of dimensionality, it is not an easy task to find these structures. The purpose of this chapter is to present a statistical approach to model the functional relationship between genes and also between genes and disease status, with special focus on nonlinear relationships. In doing so, we also identify (select), for classification purposes, the genes which are significantly more influential than the others. In data sets that we have investigated, out method shows equal ability to classify but uses far fewer genes to do so.

One of the key goals of microarray data is to perform classification via different expression profiles. In principle, gene expression profiles might serve as molecular fingerprints that would allow for accurate classification of diseases. The underlying assumption is that samples from the same class share expression profile patterns unique to their class (Yeang et al. 2001). In addition, these molecular fingerprints might reveal newer taxonomies that previously have not been readily appreciated.

Several studies have used microarrays to profile colon, breast and other tumors and have demonstrated the potential power of expression profiling for classification (Alon et al. 1999; Hedenfalk et al. 2001). Such problems can be classified as unsupervised, when only the expression data are available, and supervised, when a response measurement is also taken for each sample. In unsupervised problems (clustering) the goal is mainly to identify distinct sets of genes with similar expression profiles, suggesting that they may be biologically related. Both supervised and unsupervised problems also focus on finding sets of genes that relate to different kinds of diseases, so that future samples can be classified correctly. Classical statistical methods for clustering and classification have been applied extensively to microarray data, see Eisen et al. (1998) and Alizadeh et al. (2000) for clustering and Golub et al. (1999) and Hedenfalk et al. (2001) for classification. One of the objectives of this study is to identify sets of significant genes for classification, i.e., variable selection.

A common objective in microarray studies is to highlight genes that (on average) co-regulate with tissue type. This can be treated within a classification framework, where the tissue type is the response and the gene expressions are predictors. In this chapter we will consider rule based classifiers to discover genes that co-regulate and hence provide some of most explicit representations of the classification scheme. Rule based classifiers use primitives such as IF A THEN B, where A relates to conditions on the value of a set of predictors (genes) $\mathbf{X}$ and consequence B relates to change in $\Pr(\mathbf{Y}|\mathbf{X})$. These type of rules are easy to interpret. The best known such models are Classification and Regression Trees (CART; Breiman et al. 1984), where decision trees provide a graphical order of the rules. The objective of this chapter is two-fold: (1) find significant genes of interest and (2) find the underlying nonlinear functional form of the gene interaction. Related approaches in literature such as Lee et al. (2003) consider only linear functions of the genes, which may not be able to model such complex functional forms.

In this chapter we propose to use unordered rule sets based on a Bayesian nonparametric regression approach to model the high dimensional gene expression data. In order to explore the complex nonlinear form of the expected responses without knowledge about the functional form in advance, it is imperative that we look to nonparametric techniques, since parametric models will not be flexible enough to model these complex functions. To capture the linear dependencies, and perhaps more crucially the nonlinear functional structures between the genes we use a Bayesian version of Multivariate Adaptive Regression Splines (MARS), proposed by Friedman (1991) and extended in the Bayesian framework (BMARS) by Denison et al. (1998b). MARS is a popular method for flexible regression modeling of high dimensional data and has been extended to deal with classification problems, see for example Kooperberg et al. (1997).

In this chapter we treat the classification problem in a logistic regression framework. The logistic link has a direct interpretation of the log odds of having the disease in terms of the explanatory variables (genes). Since our model space is very large, i.e. with p genes we have 2^p models, exhaustive computation over this model space is not possible. Hence Markov Chain Monte Carlo (MCMC; Gilks et al. 1996) based stochastic search algorithms are used. Our approach is to identify significant set(s) of genes over this vast model space, first to classify accurately and then to model the functional relationship between them. The flexible nonparametric setup creates a powerful predictive model, but unlike many black box predictive machines, our method identifies the significant genes as well as focuses on the interactions among them. In this sense, the method has the advantage that it combines scientific interpretation with accurate prediction.

In order to illustrate our methodology, we choose as examples two publicly available data sets: Leukemia Data (Golub et al. 1999) and Hereditary Breast Cancer data (Hedenfalk et al. 2001). For each case we find sets of genes that have discriminating power. We also find the functional form of the main effect of dominant genes and the interaction function between genes that have significant interactions.

3.2 Model Formulation

For a binary class problem the response is usually coded as $Y_i = 1$ for class 1 and $Y_i = 0$ for the other class, where $i = 1, \ldots, n$ and where n is the number of samples (arrays). Gene expression data for p genes for n samples is summarized in an $n \times p$ matrix, $\mathbf{X}$, where each element x_{ij} denote the expression level (gene-expression value) of the jth gene in the ith sample where $j = 1, \ldots, p$. The exact meaning of expression values may be different for different matrices, representing absolute or comparative measurements, see Brazma et al. (2001). Our objective is to use the training data $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ to estimate $p(\mathbf{X}) = \Pr(\mathbf{Y} = 1|\mathbf{X})$ or alternatively the logit function $f(\mathbf{X}) = \log[p(\mathbf{X})/(1 - p(\mathbf{X}))]$.

Assume that the Y_i's are independent Bernoulli random variables with $\Pr(Y_i = 1) = p_i$ so that, $p(Y_i|p_i) = p_i^{Y_i}(1 - p_i)^{1-Y_i}$. We construct a hierarchical Bayesian model for classification as thus. Writing $p_i = \exp(\omega_i)/[1 + \exp(\omega_i)]$, wherein ω_i's are the latent variables introduced in the model to make Y_i's conditionally independent given the ω_i's. We relate ω_i to $f(\mathbf{X}_i)$ as,

$$\omega_i = f(\mathbf{X}_i) + \epsilon_i, \tag{3.1}$$

where $\mathbf{X}_i$ is the ith row of the gene expression data matrix $\mathbf{X}$ (vector of gene expression levels of the ith sample) and ϵ_i are residual random effects. The residual random effects account for the unexplained sources of variation in the data, most probably due to explanatory variables (genes) not included in the study.

We choose to model f in nonparametric framework, primarily due to the fact that parametric approaches are not flexible enough to model such "rich" gene expression datasets. One of the most common choices for f is to use a basis function method of the form,

$$f(\mathbf{X}_i) = \sum_{i=1}^{k} \beta_j B(\mathbf{X}_i, \theta_j),$$

where $\boldsymbol{\beta}$ are the regression coefficients for the bases $B(\mathbf{X}_i, \theta_j)$, which are non-linear functions of $\mathbf{X}_i$ and θ. Examples of basis function include regression splines, wavelets, artificial neural networks and radial bases. We choose a MARS basis function proposed by Friedman (1991) to model f as,

$$f(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \prod_{l=1}^{z_j} (x_{id_{jl}} - \theta_{jl})_{q_{jl}}, \tag{3.2}$$

where k is the number of spline basis, $\beta = \{\beta_1, \ldots, \beta_k\}$ are the set of spline coefficients (or output weights), z_j is the interaction level (or order) of the jth spline, θ_{jl} is a spline knot point, d_{jl} indicates which of the p predictors (genes) enters into the lth interaction of the jth spline, $d_{jl} \in \{1, \ldots, p\}$, and q_{jl} determines the orientation of the spline components, $q_{jl} \in \{+, -\}$ where $(a)_+ = \max(a, 0)$, $(a)_- = \min(a, 0)$. We choose the MARS basis function as it can flexibly model the functional relationship between explanatory variables (genes) and gives interpretable models as compared to black box techniques such as artificial neural networks.

We illustrate this rather complex notation (3.2) through an example. Suppose a MARS model is of the following form (dropping the subscript i),

$$f = 2.5 + 3.2(x_{20} - 2.5)_+ + 4.1(x_{10} - 1.2)_-(x_{30} + 3.4)_+$$

Here we have $k = 2$ spline basis functions with $\beta = \{2.5, 3.2, 4.1\}$ as the spline coefficients. Gene 20 enters the model as a linear term (main effect) with interaction level $z_1 = 1$, knot point $\theta_{11} = 2.5$, spline orientation $q_{11} = +$. We observe a bivariate interaction between genes 10 and 30 i.e. $d_{21} = 10, d_{22} = 30$ with corresponding knots $= (1.2, -3.4)$ and spline orientation $= (-, +)$. See Friedman (1991) for a comprehensive illustration of the model.

Write (3.1) and (3.2) in matrix form as,

$$\boldsymbol{\omega} \;=\; \Theta\beta + \epsilon, \tag{3.3}$$

where $\boldsymbol{\omega}$ is the vector of the latent variables, and Θ is the MARS basis matrix,

$$\Theta = \begin{bmatrix} 1 & \prod_{l=1}^{z_1}(x_{1d_{1l}} - \theta_{1l})_{q_{1l}} & \cdots & \prod_{l=1}^{z_k}(x_{1d_{kl}} - \theta_{kl})_{q_{kl}} \\ 1 & \prod_{l=1}^{z_1}(x_{2d_{1l}} - \theta_{1l})_{q_{1l}} & \cdots & \prod_{l=1}^{z_k}(x_{2d_{kl}} - \theta_{kl})_{q_{kl}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \prod_{l=1}^{z_1}(x_{nd_{1l}} - \theta_{1l})_{q_{1l}} & \cdots & \prod_{l=1}^{z_k}(x_{nd_{kl}} - \theta_{kl})_{q_{kl}} \end{bmatrix} \tag{3.4}$$

In order to aid a Bayesian formulation we impose a prior structure on all the model parameters, $\mathcal{M} = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{q}, \boldsymbol{d}, \boldsymbol{z}, v, k, , \boldsymbol{\lambda}, \sigma^2\}$. The specific forms of the priors that we take are as follows. We assign a Gaussian prior to $\boldsymbol{\beta}$ with mean $\mathbf{0}$ and variance $\sigma^2\boldsymbol{D}^{-1}$, where $\boldsymbol{D} \equiv \text{diag}(\lambda_1, \lambda, \ldots, \lambda)$ is $(n+1) \times (n+1)$ diagonal matrix. We fix λ_1 to a small value, amounting to a large variance for the intercept term but keep λ unknown. We assign a Inverse-Gamma(IG) prior to σ^2 and a gamma prior to λ with parameters (γ_1, γ_2) and (τ_1, τ_2) respectively. Note that the above model can be extended to have multiple prior variances on $\boldsymbol{\beta}$ as,

$$p(\boldsymbol{\beta}, \sigma) \sim \text{N}_{n+1}(\boldsymbol{\beta}|0, \sigma^2\boldsymbol{D}^{-1})\text{IG}(\sigma^2|\gamma_1, \gamma_2)$$

where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{n+1})^T$. Once again λ_1 is fixed to a small value but all other λ's are unknown. We assign independent Gamma (τ_1, τ_2) priors to them.

The prior structure on the MARS model parameters are as follows. The prior on the individual knot selections θ_{jl} is taken to be uniform over the n data points $p(\theta_{jl}|d_{jl}) = U(x_{1d_{jl}}, x_{2d_{jl}}, \ldots, x_{nd_{jl}})$, where d_{jl} indicates which of the genes enter our model and $p(d_{jl})$ is uniform over the p genes, $p(d_{jl}) = U(1, \ldots, p)$. The prior on the orientation of the spline is again uniform, $p(q_{jl} = +) = p(q_{jl} = -) = 0.5$. The interaction level in each spline has a prior, $p(z_j) = U(1, \ldots, z_{\text{max}})$, where z_{max} is the maximum level of interaction set by the user. Finally the prior on k, the number of splines, is taken to an improper one, $p(k) = U(1, \ldots, \infty)$, which indicates no a priori knowledge on the number of splines. Hence the model now has only one user defined parameter, z_{max}, the maximum level of interaction, for which we shall recommend a default setting in Section 3.5.

3.3 Computation

The information from the data are combined with the prior distributions on the parameters via Bayes' theorem and the likelihood function as,

$$\begin{aligned} p(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{q}, \boldsymbol{d}, \boldsymbol{z}, \boldsymbol{\beta}, v, k, \boldsymbol{\lambda}, \sigma^2|\mathbf{Y}) &= p(\mathbf{Y}|\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{q}, \boldsymbol{d}, \boldsymbol{z}, \boldsymbol{\beta}, v, k, \boldsymbol{\lambda}, \sigma^2) \\ &\quad \times p(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{q}, \boldsymbol{d}, \boldsymbol{z}, \boldsymbol{\beta}, v, k, \boldsymbol{\lambda}, \sigma^2) \end{aligned}$$

For classification problems with binary data and logistic likelihood, conjugate priors do not exist for the regression coefficients. With the Bayesian hierarchical structure as in the previous section the posterior distributions are not available in explicit form, so we use MCMC techniques (Gilks et al. 1996) for inference. Conventional MCMC methods such as the Metropolis-Hastings (MH) algorithm (Metropolis et al. 1953; Hastings 1970) are not applicable here since the parameter (model) space is variable: we do not know the number of splines apriori. Hence we use the variable dimension reversible jump algorithm outlined in Green (1995).

In our framework, the chain in updated using the following proposals with equal probability:

1. Add a new spline basis to the model.

2. Remove one of the k existing spline bases from the model.

3. Alter an existing spline basis in the model (by changing the knot points).

Following each move an update is made to the spline coefficients $\boldsymbol{\beta}$. Note that the above three move steps are equivalent to adding, removing and altering a column of Θ in (3.4). The algorithm is included in the Appendix B. The update to $\boldsymbol{\beta}$ is the critical step determining the efficiency of the algorithm. A poor proposal distribution for $\boldsymbol{\beta}$ results in the current state having low posterior probabilities and low acceptance rates. This is because adding, deleting or altering a column of Θ in (3.4) would alter the remaining $\boldsymbol{\beta}$ parameters as they are now ill-tuned to the data.

We introduce the latent variables $\boldsymbol{\omega}$ to circumvent the problem. The idea is to introduce an extra set of parameters into the model that leave the original (marginal) model distribution unchanged, in order to improve the overall efficiency of the sampling algorithms. Therefore conditional on $\boldsymbol{\omega}$, all the other parameters are independent of $\mathbf{Y}$. This allows us to adopt conjugate priors for $(\boldsymbol{\beta}, \sigma^2)$ to perform the MCMC calculations as well as marginalize over the model space. Considerable computational advantage is gained from the fact that the posterior distribution of $\boldsymbol{\beta}$ given the other parameters is now known exactly, i.e., normally distributed. The details of the procedure are given in Appendix .

3.4 Prediction and Model Choice

For a new sample with gene expression $\boldsymbol{x}_{new}$, the marginal posterior distribution of the new disease state, y_{new} is given by

$$\Pr(y_{new} = 1|\boldsymbol{x}_{new}) = \sum_{k=1}^{\infty} \int P(y_{new} = 1|\boldsymbol{x}_{new}, \mathcal{M}_k) P(\mathcal{M}_k|Y) d\mathcal{M}_k, \qquad (3.5)$$

where $\Pr(\mathcal{M}_k|Y)$ is the posterior probability and $\mathcal{M}_k$ indicates the MARS model with k splines. The integral given in (3.5) is computationally and analytically intractable and needs approximate procedures. We approximate (3.5) by its Monte Carlo estimate by,

$$\Pr(y_{new} = 1|\boldsymbol{x}_{new}) = \frac{1}{m} \sum_{j=1}^{m} P(y_{new} = 1|\boldsymbol{x}_{new}, \mathcal{M}^{(j)}), \qquad (3.6)$$

where $\mathcal{M}^{(j)}$ for $j = 1, \ldots, m$ are the m MCMC posterior samples of the MARS model parameters $\mathcal{M}$. The approximation (3.6) converges to the true value (3.5) as $m \rightarrow \infty$.

In order to select from different models, we generally use misclassification error. When a test set is provided, we first obtain the posterior distribution of the parameters based on training data, y_{trn} (train the model) and use them to classify the test samples. For a new observation from the test set, $y_{i,\text{test}}$ we will obtain the probability $\Pr(y_{i,test} = 1|y_{trn}, x_{i,test})$ by using the approximation to (3.5) given by (3.6). When this probability is greater than 0.5 we will classify it as 1 and when it is less than 0.5 we will classify it as 0. The number of misclassified samples from the test set is defined as the misclassification error.

If there is no test set available, we will use a hold-one-out cross-validation approach. For the cross validation predictive density, in general, let $\mathbf{Y}_{-i}$ be the vector of Y_j's without the ith observation Y_i,

$$P(Y_i|\mathbf{Y}_{-i}) \;=\; \tfrac{P(\mathbf{Y})}{P(\mathbf{Y}_{-i})} \;=\; \left[\int \{P(y_i|\mathbf{Y}_{-i}, \mathcal{M}_k)\}^{-1} P(\mathcal{M}_k|\mathbf{Y}) d\mathcal{M}_k\right]^{-1}.$$

The MCMC approximation to this is

$$\widehat{P}(Y_i|\mathbf{Y}_{-i,trn}) = m^{-1}\sum_{j=1}^{m}\left\{P(y_i|\mathbf{Y}_{-i,trn},\mathcal{M}^{(j)})\right\}^{-1},$$

where $\mathcal{M}^{(j)}$ for $j = 1,\ldots,m$ are the m MCMC posterior samples of the MARS model parameters $\mathcal{M}$. This simple expression is due to the fact that the Y_i's are conditionally independent given the model parameters $\mathcal{M}$.

3.5 Examples

We illustrate the Bayesian methodology with two microarray examples. For all the examples considered below we set the maximum level of interaction, $z_{\text{max}} = 2$, i.e., allow for only additive and bivariate interactions. The MCMC chain is run for 50,000 iterations of which the first 10,000 are discarded as burn-in.

3.5.1 Leukemia Data

This microarray data set is taken from Golub et al. (1999). The data set contains measurements corresponding to samples from Bone Marrow and Peripheral blood samples taken from 72 patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). As in the original paper we split the data into a training set of 38 samples (27 are ALL and 11 AML) and a test set of 34 samples (20 ALL and 14 AML). The data set contains expression levels for 7129 human genes produced by Affymetrix high-density oligonucleotide microarrays.

In order to identify significant genes, we isolate those genes that enter our MARS model most frequently in the posterior samples. Table 3 shows the genes that occur most frequently as main effects in our model. The other genes' main effects were not observed frequently in the generated MCMC samples so did not have great influence on the response. The corresponding plots of the posterior mean main effect functions

Table 3. Leukemia data: Top 50 genes (predictors) entering the Bayesian MARS model as main effects ranked in descending order of the frequency of times they appear in posterior MCMC samples

Gene ID	Gene description	Frequency
X95735	Zyxin	1.206
J04615	SNRPN Small nuclear ribonucleoprotein polypeptide N	1.018
M62762	ATP6C Vacuolar H+ ATPase proton channel subunit	0.888
J04027	Adenosine triphosphatase mRNA	0.752
X64364	BSG Basigin	0.388
Z11793	Selenoprotein P	0.354
U29091	GB DEF = Selenium-binding protein (hSBP) mRNA	0.352
U26710	Cbl-b mRNA	0.32
Z17240	HMG2 High-mobility group (nonhistone chromosomal) protein 2	0.32
L00058	MYC V-myc avian myelocytomatosis viral oncogene homolog	0.266
U79285	GLYCYLPEPTIDE N-TETRADECANOYLTRANSFERASE	0.19
X62320	GRN Granulin	0.186
U10323	Nuclear factor NF45 mRNA	0.18
M80254	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE	0.172
HG2280-HT2376	D-Amino-Acid Oxidase	0.172
M15059	”FCER2 Fc fragment of IgE, low affinity II, receptor for (CD23A)”	0.166
L07956	”GBE1 Glucan (1,4-alpha-), branching enzyme 1 (glycogen branching enzyme, Andersen disease, glycogen storage disease type IV)”	0.146
U02388	”LTB4H Leukotriene B4 omega hydroxylase (cytochrome P450, subfamily IVF)”	0.14
M19888	SPRR1B Small proline-rich protein 1B (cornifin)	0.138
M15841	SNRPB2 Small nuclear ribonucleoprotein polypeptide B”	0.138
X66363	SERINE/THREONINE-PROTEIN KINASE PCTAIRE-1	0.124
U90919	Clones 23667 and 23775 zinc finger protein mRNA	0.122
S73885	TFAP4 Transcription factor AP-4 (activating enhancer-binding protein 4))	0.118
Y12812	RFXAP mRNA	0.112
U97188	Putative RNA binding protein KOC (koc) mRNA	0.104
U82759	GB DEF = Homeodomain protein HoxA9 mRNA	0.086
X59417	PROTEASOME IOTA CHAIN	0.084
U61849	NPTX1 Neuronal pentraxin I	0.082
HG2604-HT2700	Pan-2	0.082
X60655	EVX1 Even-skipped homeo box 1 (homolog of Drosophila)	0.082
X59131	D13S106 mRNA for a highly charged amino acid sequene	0.08
L40400	”(clone zap113) mRNA, 3’ end of cds”	0.08
Z33642	V7 mRNA for leukocyte surface protein	0.08
X74570	Gal-beta(1-3/1-4)GlcNAc alpha-2.3-sialyltransferase	0.078
M64571	MAP4 Microtubule-associated protein 4	0.078
X99585	SMT3B protein	0.076
L40393	(clone S171) mRNA	0.074
L12760	”PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC”	0.072
L00022	IG EPSILON CHAIN C REGION	0.07
J00209	”IFNA10 Interferon, alpha 10”	0.066
U93205	Nuclear chloride ion channel protein (NCC27) mRNA	0.062
D63880	KIAA0159 gene	0.06
L36818	INPPL1 Inositol polyphosphate phosphatase-like protein 1 (51C protein)	0.06
HG2743-HT2846	”Caldesmon 1, Alt. Splice 4, Non-Muscle”	0.06
X92110	HcgVIII protein	0.06
X05997	GB DEF = Gastric lipase	0.058
HG4185-HT4455	”Estrogen Sulfotransferase, Ste”	0.058
L78132	Prostate carcinoma tumor antigen (pcta-1) mRNA	0.058
L40586	IDS Iduronate 2-sulfatase (Hunter syndrome)	0.056
D26067	”KIAA0033 gene, partial cds”	0.056

is shown in Figure 6. These curves are estimated by,

$$E\{f_i(X)\} = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{j:z_j=1\\ d_{jl}=i}} \beta_j^{(t)}\Theta_j^{(t)}(X), \tag{3.7}$$

where T is the number of models in the generated sample, indexed with the superscript. The second summation ensures that the curves are estimated by only considering the main effect basis functions involving the i gene (predictor), thus averaging over the basis functions relating to the desired gene main effect. Note here that these curves can only guide us to the shape of the main effect functions. We can see that there is evidence that gene BSG Basigin shows little effect on the response. As the expression level of gene Adenosine triphosphatase mRNA increases the response decreases linearly. Observing the plots of main effect functions (Figure 6), for the gene Zyxin the response is unaffected over the negative expression values but decreases linearly for increasing positive expression values, while on the other hand exactly the opposite feature is found gene SNRPN Small nuclear ribonucleoprotein polypeptide N where the response increases linearly for negative expression values and is unaffected in the positive range. Similar conclusions can be drawn for other genes too. This demonstrates how threshold basis functions such as MARS, allow for insightful interpretation of the relationship between response and genes (predictors), with the added advantage being that MARS model automatically ignores variables that have little effect on the response.

Table 4 shows the genes that enter as an interaction term most often in the posterior samples. Figure 7 shows the interaction surface of the top three interacting gene pairs indicating the joint contribution to the odds of having a disease of the two genes. The surface is estimated in a manner similar to (3.7), but now we only consider interaction terms involving the two genes desired in the second summation. This figure highlights the advantage of the using flexible nonlinear MARS

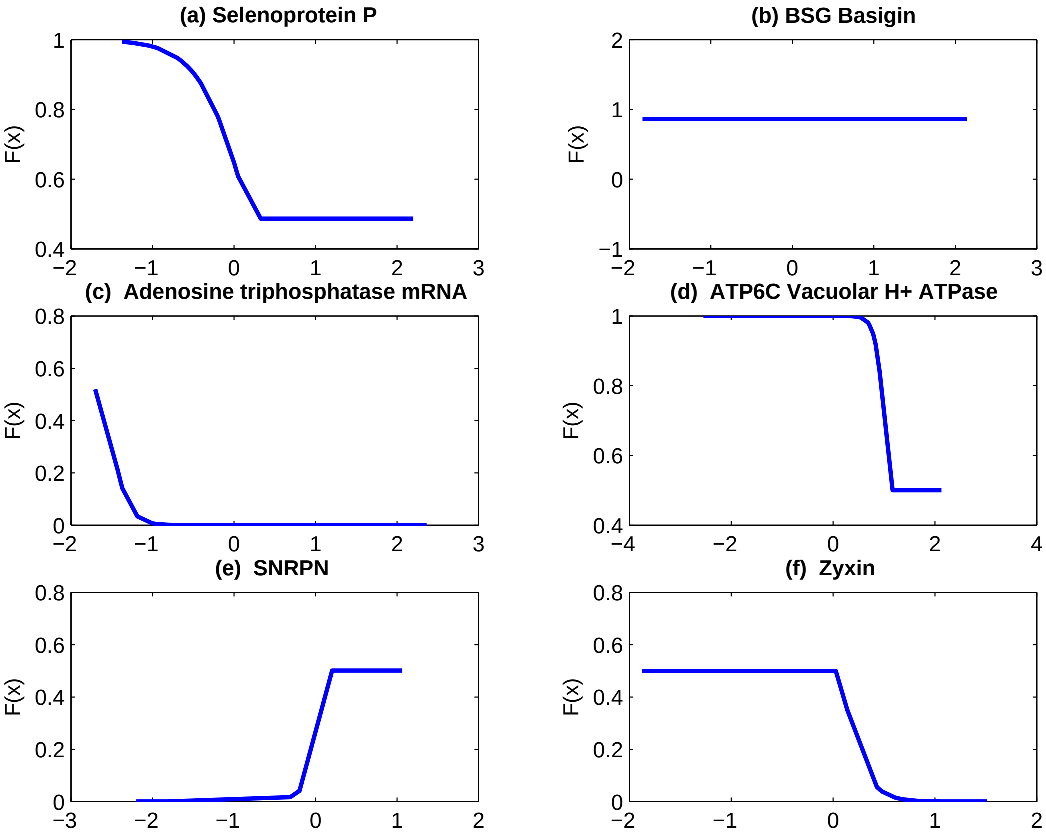


Figure 6. Leukemia data: Posterior mean main effects of the significant genes entering the Bayesian MARS model. The horizontal axis is the standardized expression level of the gene and the vertical axis is the mean main effect function.

basis functions in discerning this complex interaction function between genes over linear approaches. In the top panel of Figure 7 we can see that high expression levels of gene Alpha-Amylase 2B Precursor combined with low (negative) expression levels of gene Adenosine triphosphatase calcium results in an increased level of response, which is unaffected for low levels of both genes. A similar feature is also detected observing the interaction surface of genes Natural killer cell receptor (KIR) mRNA and HOXA1 Homeo box A1. From the bottom most panel we can observe that the odds increases as the expression level of gene LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog increases and that of gene DAGK1 Diacylglycerol kinase alpha decreases.

Using two pairs of genes that discriminate between the two classes AML and ALL reasonably, we plot the probability contours, $\Pr(Y_i = 1)$ (Figure 8) to demonstrate the advantages of using a nonlinear model. Any linear approach would divide the predictor space into two regions separated by a straight line. Such a complex decision boundary can only be uncovered using a nonlinear model. Note that this predictive contours appear smooth even though individual MARS models have axis parallel non-smooth contours. This is due to averaging over thousands of MARS models, thus marginalizing over the model space.

Golub et al. use a 50-gene predictor trained using their weighted voting scheme on the training samples. The predictor made strong predictions for 29 of the 34 test samples, declining to predict the other five cases. For the same case our misclassification error rate for the test set is 0.08, i.e, we misclassify 3 out of the 34 test samples. Our results appear to be competitive to the results from Golub et al., but we use far fewer genes. Figure 9(a) shows the marginal density of the number of splines, $p(k|Y)$ from 50,000 samples of our MCMC chain. The mode of the distribution is 2 basis terms, thus showing that we get competitive results by using considerably fewer

Table 4 : *Leukemia data: Top 25 interacting genes entering the Bayesian MARS model ranked in descending order of the frequency of times they appear in posterior MCMC samples*

Image Clone ID	Gene description	Image Clone ID	Gene description	Frequency
D90097	ALPHA-AMYLASE 2B PRECURSOR	Z69881	"Adenosine triphosphatase, calcium"	1.52
U37431	HOXA1 Homeo box A1	X99479	Natural killer cell receptor (KIR) mRNA	0.634
M16038	LYN V-yes-1 Yamaguchi sarcoma viral	X62535	"DAGK1 Diacylglycerol kinase, alpha (80kD)"	0.602
U35451	Heterochromatin protein p25 mRNA	X78712	GKP2 Glycerol kinase 2 (testis specific)	0.526
M77142	NUCLEOLYSIN TIA-1	M92934	CTGF Connective tissue growth factor	0.25
S74728	Antiquitin	X13589	"CYP19 Cytochrome P450, subfamily XIX"	0.246
HG511-HT511	Ras Inhibitor Inf	U25956	SELPLG Selectin P ligand	0.23
L25085	PROTEIN TRANSPORT PROTEIN SEC61	X89101	"GB DEF = Fas (Apo-1, CD95)"	0.224
AF00893	GB DEF = Syntaxin-16C mRNA	U90552	Butyrophilin (BTF5) mRNA	0.212
Z26876	LTBP1 Latent transforming growth factor	M83652	"PFC Properdin P factor, complement"	0.168
HG1612-HT1612	Macmarcks	L20971	"PDE4B Phosphodiesterase 4B, cAMP-specific "	0.164
M31013	"MYH9 Myosin, heavy polypeptide 9"	HG982-HT982	Pre-T/Nk-Cell-Associated Protein 1f6	0.164
M14764	NGFR Nerve growth factor receptor	U57796	Zinc finger protein (LD5-1) mRNA	0.16
D28383	"GB DEF = ATP synthase B chain, 5'UTR "	L39009	"GB DEF = Class IV alcohol dehydrogenase 7 (ADH7)"	0.152
U39400	NOF1 mRNA	U78313	Myogenic repressor I-mf (MDFI) mRNA	0.15
D63481	"KIAA0147 gene, partial cds"	M86707	GLYCYLPEPTIDE N-TETRADECANOYLTRANSFERASE	0.146
D38555	KIAA0079 gene	U44975	"DNA-binding protein CPBP (CPBP) mRNA, partial cds"	0.124
D14889	"Small GTP-binding protein, S10"	L23116	GALC Galactocerebrosidase	0.094
M36653	"POU2F2 POU domain, class 2"	X00351	"ACTB Actin, beta"	0.094
M33478	PHOSDUCIN	M24766	"COL4A2 Collagen, type IV, alpha 2"	0.092
X74295	"ITGA7 Integrin, alpha 7B"	X97074	EEF2 Eukaryotic translation elongation factor 2	0.084
M92432	"GUC2D Guanylate cyclase 2D"	S82470	BB1	0.078
L36983	Dynamin (DNM) mRNA	X70811	"ADRB3 Adrenergic, beta-3-, receptor"	0.078
U48408	Kidney water channel (hKID) mRNA	U62136	"Putative enterocyte differentiation promoting factor mRNA "	0.074
D12763	ST2 Suppression of tumorigenicity 2	U21090	DNA polymerase delta small subunit mRNA	0.074

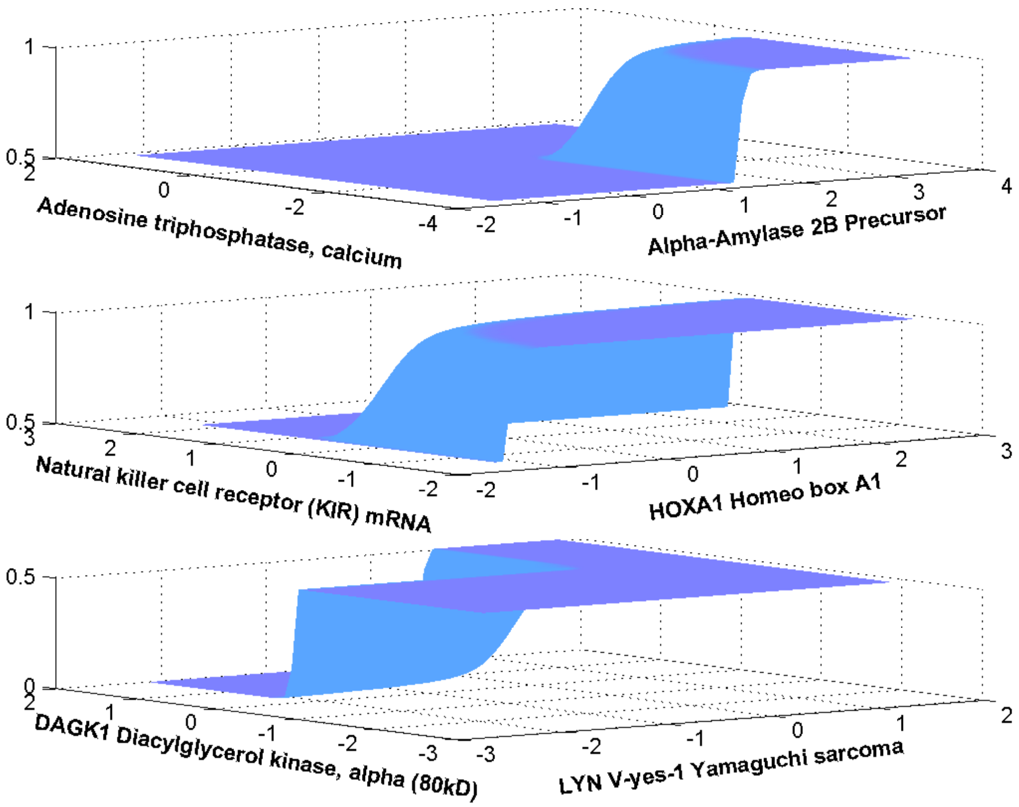


Figure 7. Leukemia data: Posterior mean interaction functions of the significant genes entering the Bayesian MARS model. The X and Y axes are the standardized expression levels of the interacting genes and the vertical axis is the mean interaction function.

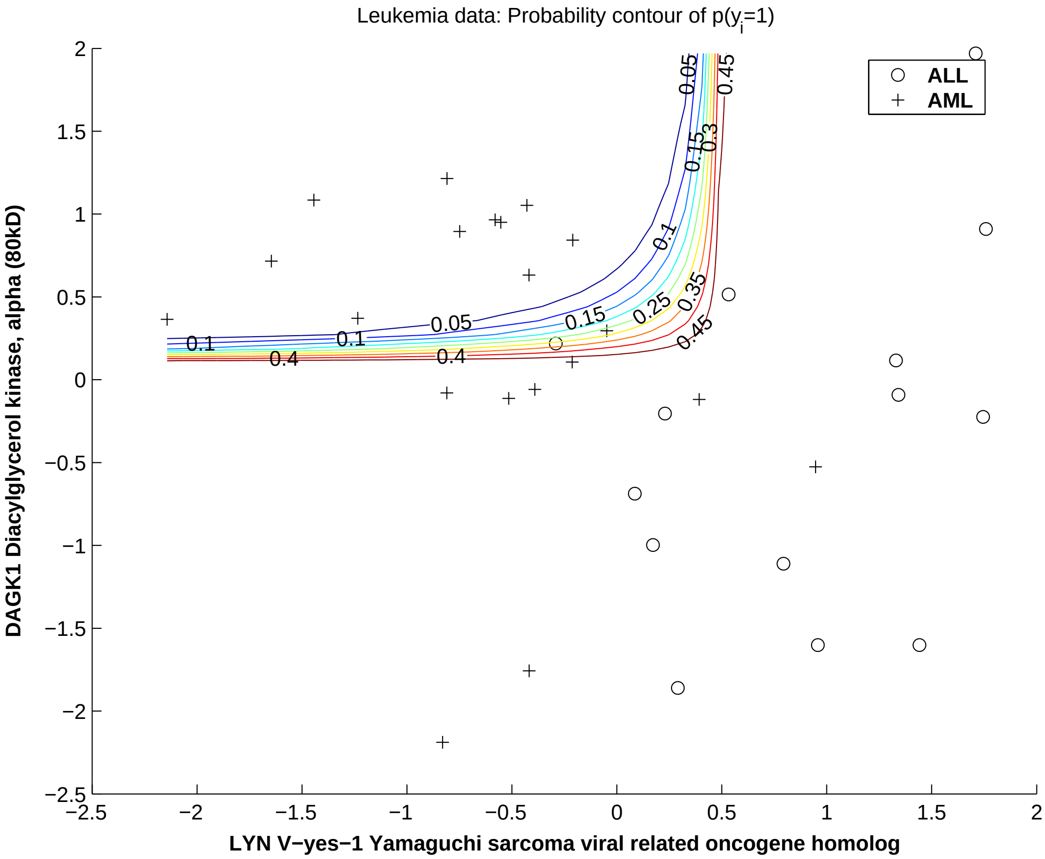


Figure 8. Leukemia data: Probability contours showing $P(Y_i = 1)$ under the Bayesian MARS model. The circles and the crosses represent diseased and non-diseased respectively.

genes.

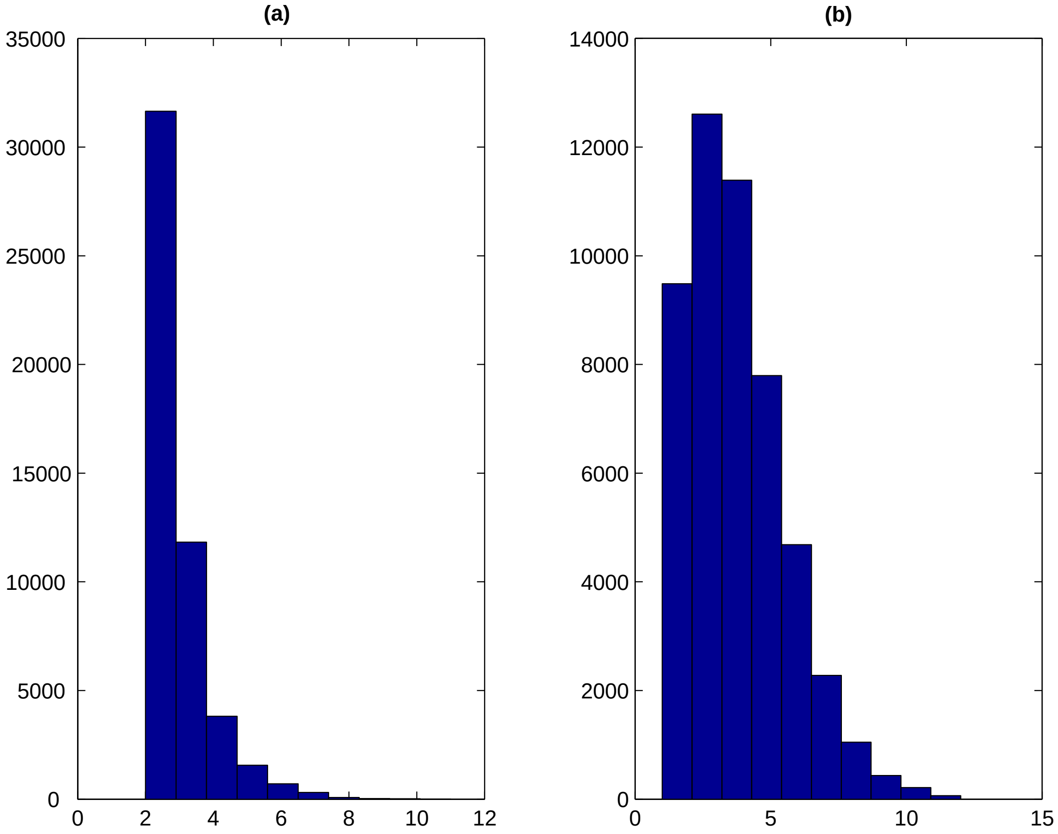


Figure 9. Plot of $p(k|Y)$, the posterior distribution of the number of splines basis functions using 50,000 samples from the MCMC output. (a) Leukemia data; (b) Breast cancer data.

3.5.2 Hereditary Breast Cancer Data

We use the microarray data-set used in Hedenfalk et al. (2001) on breast tumors from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2 or from patients not expected to carry a hereditary predisposing mutation. Pathological and genetic differences appear to imply different but overlapping functions for BRCA1 and BRCA2. They examined 22 breast tumor samples from 21 breast cancer patients, and all patients except one were women. Fifteen women had hereditary breast cancer, 7 tumors with BRCA1 and 8 tumors with BRCA2. 3,226 genes were used for each breast tumor sample. We use our method to classify BRCA1 versus the others (BRCA2 and sporadic).

Table 5 lists the top 50 genes that enter as main effects in the MARS model in posterior MCMC samples, along with the corresponding frequency of appearance. Similarly Table 6 shows the top 25 interacting genes that enter the MARS model. These genes enter our model most frequently while classifying BRCA1 versus BRCA2 and sporadic. A similar list of 51 genes which best differentiate among the types of tumor is also provided by Hedenfalk et al.. We find quote a few overlapping genes (marked by a *) between the two lists like keratin 8 (KRT8), ODCantizyme and ACTR1A. KRT8 is a member of the cytokeratin family of genes and cytokeratins are frequently used to indentify breast cancer metastases by immunohistochemistry, and cytokeratin 8 abundance has been shown to correlate well with node-positive disease (Brotherick et al. 1998).

Figure 10 shows the posterior mean main effect function of the top six genes genes selected from the list. The vertical axis show the odds of having BRCA1 mutation and the horizontal axis is the standardized expression level of that particular gene. An advantage of using a nonlinear approach is evident here as we can unearth a threshold expression level and its corresponding effect on the the odds of having a BRCA1 mutation. For example, for polymerase (RNA) II polypeptide it is seen that the odds are relatively high for negative expression levels while the odds decrease for higher expression levels of the gene. Figure 11 shows posterior mean interaction function of two pairs of genes from the list of top 25. This shows the combined effect these two genes on the odds of carrying mutation of BRCA1.

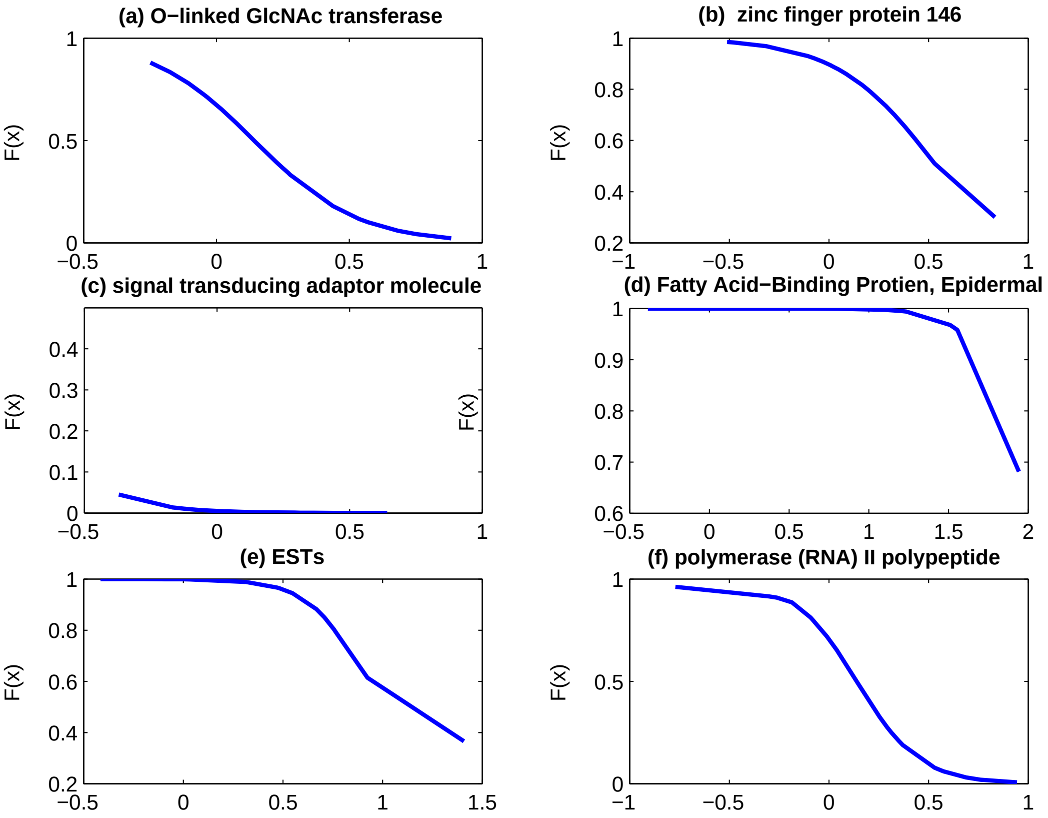
Since test data were not provided, to check our model adequacy we used full hold-one-out cross validation. The results are summarized in Table 7. We compare our cross validation results with other popular classification algorithms as in Lee et

Table 5. Breast cancer data: Top 50 genes (predictors) entering the Bayesian MARS model as main effects ranked in descending order of the frequency of times they appear in posterior MCMC samples

Image Clone ID	Gene description	Frequency
767817	polymerase (RNA) II (DNA directed) polypeptide F	3.596
307843	ESTs (*)	2.87
81331	”FATTY ACID-BINDING PROTEIN, EPIDERMAL”	2.46
843076	signal transducing adaptor molecule (SH3 domain and ITAM motif) 1	2.396
825478	zinc finger protein 146	2.304
28012	O-linked N-acetylglucosamine (GlcNAc) transferase	2.17
812227	”solute carrier family 9 (sodium/hydrogen exchanger), isoform 1 ”	2.024
566887	heterochromatin-like protein 1 (*)	1.946
841617	ornithine decarboxylase antizyme 1 (*)	1.894
788721	KIAA0090 protein	1.81
811930	KIAA0020 gene product	1.668
32790	”mutS (E. coli) homolog 2 (colon cancer, nonpolyposis type 1)”	1.46
784830	D123 gene product (*)	1.396
949932	nuclease sensitive element binding protein 1 (*)	1.382
26184	”phosphofructokinase, platelet” (*)	1.294
810899	CDC28 protein kinase 1	1.294
46019	minichromosome maintenance deficient (S. cerevisiae) 7 (*)	1.266
897781	keratin 8 (*)	1.192
32231	KIAA0246 protein (*)	1.084
293104	phytanoyl-CoA hydroxylase (Refsum disease) (*)	1.006
180298	protein tyrosine kinase 2 beta	0.952
47884	macrophage migration inhibitory factor (glycosylation-inhibiting factor)	0.864
137638	ESTs (*)	0.792
246749	”ESTs, Weakly similar to trg [R.norvegicus]”	0.792
233365	HP1-BP74	0.788
815530	PAK-interacting exchange factor beta	0.68
123425	”ESTs, Moderately similar to AF141326 RNA helicase HDB/DICE1 [H.sapiens]”	0.642
22230	”collagen, type V, alpha 1”	0.612
324210	sigma receptor (SR31747 binding protein 1)	0.608
824117	vaccinia related kinase 2	0.602
124405	androgen induced protein	0.594
83210	”Complement component 8, beta polypeptide”	0.592
49788	carnitine acetyltransferase	0.59
344352	ESTs	0.586
842806	cyclin-dependent kinase 4	0.568
810734	Human 1.1 kb mRNA upregulated in retinoic acid treated HL-60 neutrophilic cells	0.564
814701	”MAD2 (mitotic arrest deficient, yeast, homolog)-like 1”	0.554
36007	zinc finger protein 133 (clone pHZ-13)	0.518
110503	FOS-like antigen-1	0.492
767784	jun D proto-oncogene	0.488
486844	”gap junction protein, alpha 1, 43kD (connexin 43)”	0.486
810408	hypothetical 43.2 Kd protein	0.456
199381	vav 3 oncogene	0.446
509682	histone deacetylase 3	0.446
43021	histidyl-tRNA synthetase	0.438
212198	”tumor protein p53-binding protein, 2” (*)	0.418
840702	SELENOPHOSPHATE SYNTHETASE ; Human selenium donor protein (*)	0.402
666128	D component of complement (adipsin)	0.4
613126	ubiquitin specific protease 13 (isopeptidase T-3)	0.396
139705	ESTs	0.384

Table 6 : *Breast cancer data : Top 25 interacting genes entering the Bayesian MARS model ranked in descending order of the frequency of times they appear in posterior MCMC samples*

Image Clone ID	Gene description	Image Clone ID	Gene description	Frequency
753285	glycogenin	841617	ornithine decarboxylase antizyme 1 (*)	1.814
134748	glycine cleavage system protein H	137506	dishevelled 2 (homologous to Drosophila dsh)	1.296
126412	ring finger protein 14	282980	ESTs	1.232
784830	D123 gene product (*)	814270	polymyositis/scleroderma autoantigen 1 (75kD)	1.196
823663	fragile X mental retardation	377275	ataxia-telangiectasia group D-associated protein	1.156
50754	mitochondrial translational initiation factor 2	282980	ESTs	1.104
346117	guanylate binding protein 2, interferon-inducible	786083	ubiquitin-conjugating enzyme E2 variant 2	1.082
79898	transducin-like enhancer of split 1	204179	hypothetical protein FLJ20036	0.97
214537	replication factor C (activator 1) 1 (145kD)	246304	BTG family, member 3	0.882
307843	ESTs (*)	199624	ESTs	0.744
136730	TATA box binding protein (TBP)-associated factor	32790	mutS (E. coli) homolog 2	0.708
711959	polymerase (RNA) III (DNA directed) (62kD)	195947	ESTs, Weakly similar to [H.sapiens]	0.622
949932	nuclease sensitive element binding protein 1 (*)	38393	connective tissue growth factor	0.616
756847	suppressin (nuclear deformed epidermal)	29054	ARP1 (actin-related protein 1 (*)	0.592
814054	KIAA0040 gene product	309045	Sarcolemmal-associated protein	0.576
51740	hydroxyacyl-Coenzyme A dehydrogenase	340644	integrin, beta 8(*)	0.566
823930	actin related protein 2/3 complex	32609	laminin, alpha 4	0.518
297392	metallothionein 1L	366647	butyrate response factor 1 (EGF-response factor 1)	0.504
73531	nitrogen fixation cluster-like (*)	29063	Homo sapiens clone 23620 mRNA sequence	0.498
898123	phosphoribosylglycinamide formyltransferase	32231	KIAA0246 protein (*)	0.48
194364	RNA binding motif protein 6	21652	catenin (cadherin-associated protein), alpha 1	0.462
143227	ESTs	135381	growth arrest and DNA-damage-inducible 34	0.442
249705	Deleted in split-hand/split-foot 1 region	344352	ESTs	0.408
771173	hypothetical protein	784744	M-phase phosphoprotein 6	0.402
713647	tetraspan 3	240033	Homo sapiens mRNA; cDNA DKFZp434L162	0.4


(a) O−linked GlcNAc transferase

(b) zinc finger protein 146

(c) signal transducing adaptor molecule

(d) Fatty Acid−Binding Protien, Epidermal

(e) ESTs

(f) polymerase (RNA) II polypeptide

Figure 10. Breast cancer data: Posterior mean main effects of the significant genes entering the Bayesian MARS model. The horizontal axis is the standardized expression level of the gene and the vertical axis is the mean main effect function.

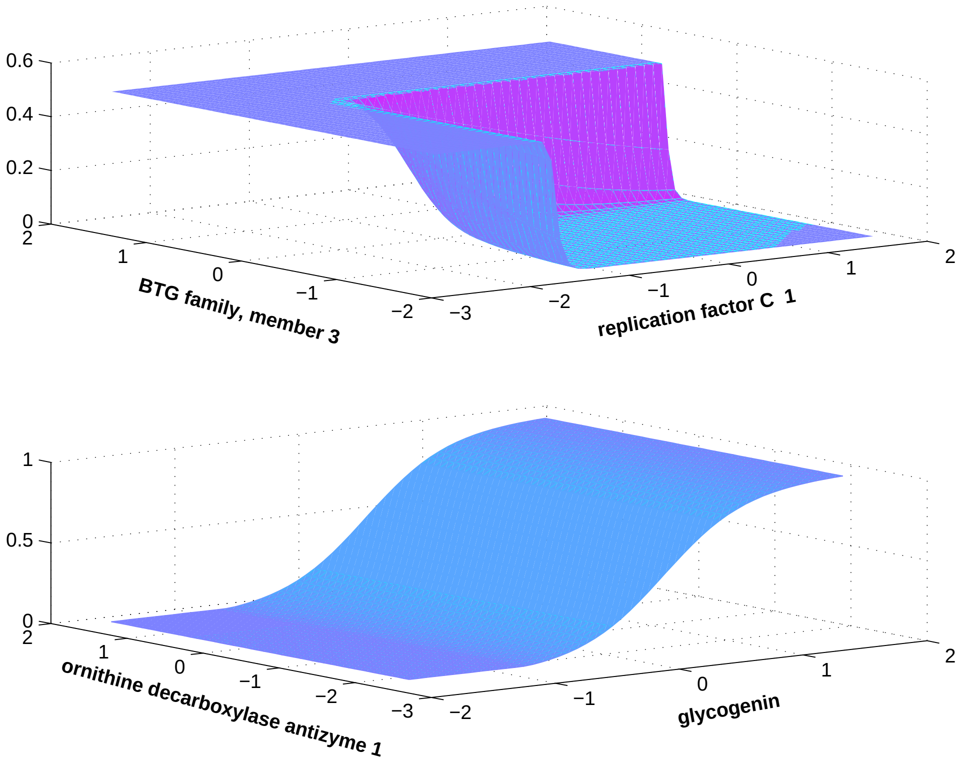


Figure 11. Breast cancer data: Posterior mean interaction functions of the significant genes entering the Bayesian MARS model. The X and Y axis are the standardized expression levels of the interact ing genes and the vertical axis is the mean interaction function.

al. (2003). All the other methods use 51 genes for classification purposes while the MARS methods selects far fewer genes. Figure 9(b) shows the marginal density of the number of splines, $p(k|Y)$ from 50,000 samples of our MCMC chain. The mode of the distribution is 3 showing that the number of splines basis terms (genes) used by the model adapts to the problem at hand and uses fewer genes with the results being competitive to any other method.

Table 7. Model misclassification errors using hold-one-out cross-validation for breast cancer data

Model	Number of misclassifies samples
Bayesian MARS	0
Feed-forward neural networks (3 hidden neurons, 1 hidden layer)	1.5 (Average error)
Gaussian kernel	1
Epanechnikov kernel	1
Moving window kernel	2
Probabilistic neural network (r=0.01)	3
kNN (k=1)	4
SVM linear	4
Perceptron	5
SVM Nonlinear	6

CHAPTER IV

SUMMARY AND FUTURE RESEARCH

In Chapter II we presented an automated Bayesian method to fit spatially adaptive curves. We also provided an extension to additive models, wherein we obtained estimates of regression function and uncertainty intervals. We use regression P-splines to model the unknown smooth function, but we allow spatial adaptivity in the penalty function by modeling it as another regression P–spline. Our simulations indicate that our method is very competitive to that of Ruppert and Carroll (2000) in terms of frequentist mean squared error and coverage probabilities of the credible intervals. We also provide Bayesian uncertainty intervals: the intervals had pointwise coverage close to the frequentist nominal level. Other methods, such as those of Donoho and Johnstone (1994) and Denison et al. (1998a) appear to be no better than ours, and in some cases worse. Simulations indicate that our methods are roughly comparable to the BARS method of DiMatteo et al. (2001).

One issue that remains unresolved is choice of the number of knots for the regression P-spline. We have here relied on the work of Ruppert (2002) in simulated data sets, the analyses of data examples in Ruppert et al. (2003), and work of Eilers and Marx (1996, 2002) as evidence that a large number of knots is unnecessary within the P-spline paradigm, with recommendations of between 10 and 50 knots for most situations of non-spatially adaptive smoothing. We believe that choosing a far smaller number of knots for spatially adaptive smoothing, as we have done, makes intuitive sense, and clearly works well in our examples, and in the examples of Ruppert and Carroll (2000). Inevitably, however, there will be interest in letting the data select the number of knots, even in the P-spline paradigm. Indeed, this has been done in the frequentist approach, see Ruppert and Carroll (2000) and Ruppert et al. (2003, Chapter 17), where the number of knots and subknots is effectively chosen by GCV. We conjecture that the following method will work in our Bayesian context: it is based on the work of Kass and Wasserman (1995) and illustrated in a model-averaging context by Carroll, Roeder and Wasserman (1999). Specifically, choose a set of numbers of knots and subknots: Ruppert, et al. use combinations of $(10, 20, 40, 80, 120)$ for the former and $(3, 4, 5, 6)$ for the latter. Then run our method for each of the combinations, and compute BIC for each at the posterior mean of median of the parameters. Finally, either select the combination on the basis of BIC, or average the fits using BIC as a model averaging device. We conjecture that this approach will work comparably to the frequentist approaches.

The other obvious issue here is the general comparison between regression spline methods. There are basically three approaches: (a) the P-spline approach as advocated here needs little more introduction; (b) knot selection methods such as BARS; and (c) regression splines without penalization but where the number of knots are selected using devices such as BIC and AIC, see Rice and Wu (2001) for an illustration. All these methods have value. Approach (c) generally tends to choose a smallish number of knots and is essentially only available in a frequentist context. Our own view is that in the frequentist context of regression splines, penalization with a fixed number of knots is a more natural approach than selecting the number of knots, especially when inference is of interest, since inference after model selection is extremely difficult in the frequentist context. The knot selection methods (free-knot splines) are clearly geared to handle problems where a high degree of spatial adaptation is necessary: that the P-spline approach does reasonably well in comparison to say BARS may in fact be seen as somewhat surprising. One nice feature of P-splines is that being little more than mixed models methods, they are readily adapted to new problems without great effort. Ciprian Crainiceanu (personal communication) has recently shown how to implement our methods in WinBUGS, which gives some sense of its ease of application.

An interesting question concerns extension of these methods to non-Gaussian data. Indeed, BARS for example was motivated originally for the treatment of Poisson data, where it is extremely effective. Bayesian versions of penalized regression splines that are not spatially adaptive are easy to develop for generalized linear models, either via brute-force (Ruppert, et al., 2003, Chapter 16) or via latent variable methods such as those Albert and Chib (1993) for binary data, and of Holmes and Mallick (2003) for binary and count data as special cases. These latter approaches essentially place one back into the Gaussian framework after the latent variables are computed, and these devices should allow spatially adaptive smoothing to be developed readily. Another interesting question is spatially adaptive smoothing in the presence of heteroscedasticity: frequentist P-splines are readily developed in this case (Ruppert, et al., 2003, chap. 14), and adding Bayesian spatially adaptive smoothing should be possible.

In Chapter II we presented a approach to model nonlinear gene interactions using a Bayesian MARS. Our method uses MCMC based stochastic search algorithms to obtain the models. The advantage of our method is that we capture the nonlinear dependencies between the genes, dependencies that would have been missed by linear approaches. Our approach is not only flexible enough to model these complex interaction functions, but it also identifies significant genes of interest for further biological study. We illustrated our method using two microarray data sets which have been well analyzed in literature. In both cases we used far fewer genes and yet obtained competitive results to those reported in literature.

We have treated the binary case in detail in this study. When the response is not binary, such that the number of classes (C) is greater than two, then the problem becomes a multiclass classification problem. This can be handled in a manner similar to the binary classification approach, as follows. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iC})$ denote the multinomial indicator vector with elements $Y_{iq} = 1$ if the qth sample is belongs to the qth class and $Y_{ij} = 0$ otherwise. Let $\mathbf{Y}$ denote the $n \times C$ matrix of these indicators. The likelihood of the data given the MARS spline bases $(\Theta_1, \ldots, \Theta_C)$, is given by,

$$\mathrm{P}(Y_i = 1|\mathbf{X}_i) = p_1^{y_{i1}} p_2^{y_{i2}}, \ldots, p_C^{y_{iC}},$$

where p_q is the probability that the sample came from class q. This is modelled in a similar manner to the binary class case as in Section 3.2. The prior structure imposed on the parameters is also akin to that described in Section 3.2. A detailed study will be performed in future.

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Alizadeh, A., Eisen, M., Davis, R. E., Chi Ma, Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., and Yu, X. et al.. (2000), "Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling," *Nature*, 403, 503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Natl. Acad. Sci.*, 96, 6745–6750.

Brazma, A., Hingamp, P., Quakenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., A., B. C., and Causton, H. C. et al.. (2001), "Minimum Information about Micorarray Experiment (MAIME) - Towards Standards for Microarray Data," *Nature Genet.*, 29, 365–371.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, California:Wadsworth Publishing Co Inc.

Brotherick, I., Robson, C. N., Browell, D. A., Shenfine, J., While, M. D., Cunliffe, W. J., Shenton, B. K., Egan, M., Webb, L. A., and Lunt, L. G. et al.. (1998), "Cytokeratin Expression in Breast Cancer: Phenotypic Changes Associated with Disease Progression," *Cytometry*, 32, 310–308.

Carroll, R. J., Roeder, K., and Wasserman, L. (1999), "Flexible Parametric Measurement Error Models," *Biometrics*, 55, 44–54.

Coull, B. A., Ruppert, D., and Wand, M. P. (2001), "Simple Incorporation of Interactions Into Additive Models," *Biometrics*, 57, 539–545.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998a), "Automatic Bayesian Curve Fitting," *Journal of Royal Statistical Society, Series B*, 60, 333–350.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998b), "Bayesian MARS," *Statistics and Computing*, 8, 337–346.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001), "Bayesian Curve Fitting with Free-knot Splines," *Biometrika*, 88, 1055–1071.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Duggan, D. J., Bittner, M. L., Chen, Y., Meltzer, P. S., and Trent, J. M. (1999), "Expression Profiling Using cDNA Microarrays," *Nature Genet.*, 21, 10–14.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties (with discussion)," *Statistical Science*, 11, 89–121.

——— (2002), "Generalized Linear Additive Smooth Structures," *Journal of Computational and Graphical Statistics*, 11, 758–783.

Eisen, M. B., Spellman., P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proc. Natl. Acad. Sci.*, 95, 14863–14868.

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines (Disc: p67-141)," *The Annals of Statistics*, 19, 1–67.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, Boca Raton, Florida: Chapman and Hall.

Golub, T. R., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek., M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caliguiri, M., and et al. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hastie, W., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., and et al. (2001), "Gene Expression Profiles in Hereditary Breast Cancer," *New England Journal of Medicine*, 344, 539–548.

Holmes, C., and Mallick, B. K. (2003), "Generalized Nonlinear Modeling with Multivariate Smoothing Splines," *Journal of the American Statistical Association*, 98, 352–368.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Kooperberg, C., Bose, S., and Stone, C. J. (1997), "Polychotomous Regression," *Journal of the American Statistical Association*, 93, 117–127.

Lee, K. Y., N., S., Doughetry, E. R., Vanucci, M., and Mallick, B. K. (2003), "Gene Selection: A Bayesian Variable Selection Approach," *Bioinformatics*, 19, 90–97.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.

Rice, J. A., and Wu, C. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253–269.

Robinson, G. K. (1991), “That BLUP Is a Good Thing: The Estimation of Random Effects (with discussion),” *Statistical Science*, 6, 15–51.

Ruppert, D. (2002), “Selecting the Number of Knots for Penalized Splines,” *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D., and Carroll, R. J. (2000), “Spatially-Adaptive Penalties for Splines Fitting,” *Australian and New Zealand Journal of Statistics*, 42 (2), 205–223.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995), “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray,” *Science*, 270, 467–470.

Wand, M. P. (2000), “A Comparison of Regression Spline Smoothing Procedures,” *Computational Statistics [Formerly: Computational Statistics Quarterly]*, 15 (4), 443–462.

Yeang, C. H., S., R., Tamayo, P., Mukherjee, S., Rifkin, R. M., Reich, M., Lander, E., Mesirov, J., and Golub, T. (2001), “Molecular Classification of Multiple Tumor Types,” *Bioinformatics*, 17, 316–322.

APPENDIX A

BAYESIAN P–SPLINES: DETAILS OF THE SAMPLER

In this section we derive the conditional distributions for the all the random variables and parameters in the spatially adaptive Bayesian P–spline model outlined in Chapter II. The algorithm for the MCMC sampling is as follows:

- Give initial values to all parameters: $\Omega_Y,\ \Omega_s,\ \xi^2,\ \{\sigma^2(\kappa_j)\}_{j=1}^{M_Y}$ and σ_Y^2.
- Start the MCMC sampler and iterate.
- *Updating* (Ω_Y, σ_Y^2)

  Conditional on the rest of the parameters, using Bayesian linear model theory with conjugate priors, the conditional posterior distribution of (Ω_Y, σ_Y^2) is,

  $$[\Omega_Y, \sigma_Y^2] \quad \sim \quad \text{Normal}(m_Y, \Sigma_Y)\text{IG}(\widetilde{a}_Y, \widetilde{b}_Y)$$

  where $m_Y = (1/\sigma_Y^2)(\Sigma_Y Z_Y^T Y)$, $\Sigma_Y = [(Z_Y^T Z_Y/\sigma_Y^2 + \Lambda_Y^{-1})^{-1}]$, Z_Y is the regression spline design matrix and $\Lambda_Y = \text{diag}\{100, \ldots, 100, \sigma^2(\kappa_1), \ldots, \sigma^2(\kappa_{M_Y})\}$ is the prior variance on Ω_Y. Here IG($\bullet$) is the Inverse Gamma distribution with shape parameter, $\widetilde{a}_Y = [(n/2) + a_Y]$ and scale parameter, $\widetilde{b}_Y = [(1/2)\{(Y - Z_Y\Omega_Y)^T(Y - Z_Y\Omega_Y)\} + (1/b_Y)]^{-1}$.
- *Updating* (Ω_s, ξ^2)

  With conjugate priors on (Ω_s, ξ^2), the conditional posterior distribution is,

  $$[\Omega_s, \xi^2] \quad \sim \quad \text{Normal}(m_s, \Sigma_s)\text{IG}(\widetilde{a}_s, \widetilde{b}_s)$$

  where $m_s = (1/\sigma_u^2)(\Sigma_s Z_s^T \rho)$, $\Sigma_s = [(1/\sigma_u^2)(Z_s^T Z_s) + \Lambda_s^{-1}]^{-1}$ and ρ denotes the vector $[-\log\{\sigma^2(\kappa_1)\}, \ldots, -\log\{\sigma^2(\kappa_{M_Y})\}]^T$. $\Lambda_s = \text{diag}\{100, 100, \xi^2, \ldots, \xi^2\}$ is the prior variance on Ω_s. The posterior inverse gamma parameters are $\widetilde{a}_s = [(M_Y/2) + a_s]$ and $\widetilde{b}_s = [(1/2)\{\sum_{j=1}^{M_s} \beta_{js}^2\} + (1/b_s)]^{-1}$. We set (a_s, b_s) to be $(1, 1)$ for all the examples.

- *Updating* $\{\sigma^2(\kappa_j)\}_{j=1}^{M_Y}$

The penalty parameters, $\{\sigma^2(\kappa_j)\}_{j=1}^{M_Y}$, conditional on the current model parameters does not have an explicit form. Thus we resort to Metropolis-Hastings procedure withe a proposal density $T[\sigma^{2*}(\kappa_j), \sigma^2(\kappa_j)]$ that generates the moves from the current state $\sigma^{2*}(\kappa_j)$ to a new state $\sigma^2(\kappa_j)$. The proposed updates are then accepted with probabilities,

$$\alpha = \min\left\{1, \frac{p[\sigma^{2*}(\kappa_j)|\text{rest}]T[\sigma^2(\kappa_j), \sigma^{2*}(\kappa_j)]}{p[\sigma^2(\kappa_j)|\text{rest}]T[\sigma^{2*}(\kappa_j), \sigma^2(\kappa_j)]}\right\},$$

otherwise the current model is retained. It is convenient to take the proposal distribution $T[\sigma^{2*}(\kappa_j), \sigma^2(\kappa_j)]$ to be a symmetric distribution (eg. Gaussian) with mean equal to the old value $\sigma^2(\kappa_j)$ and a pre-specified standard deviation. Since the density involves exponential terms, the likelihood values

calculated during the implementation of the algorithm are typically very large, hence we worked on a log scale. A common problem encountered in the implementation is the non-mobility of the MH step. If we start at bad starting values it may take a large number of iterations or even worse may not converge. To circumvent the problem we use frequentist estimates as starting values for the MCMC run, and in particular use the estimates of the smoothing parameter that minimize the generalized cross validation (GCV) statistic

$$\text{GCV} = \frac{\|Y - Z_Y\Omega_Y(\sigma^2(\kappa))\|}{[(1 - \text{df}(\sigma^2(\kappa)))/n]^2}$$

where

$$\text{df}(\sigma^2(\kappa)) = \text{tr}\{(Z_Y^T Z_Y + \Lambda_Y)^{-1}(Z_Y^T Z_Y)\}$$

is the degree of freedom of the smoother which is defined to be the trace of the smoother matrix (Hastie and Tibshirani, 1990 Section 3.5).

APPENDIX B

BAYESIAN MARS: DETAILS OF THE SAMPLER

In this section we derive the conditional distributions for the all the random variables and parameters in the Bayesian MARS model outlined in Chapter III. The algorithm for the MCMC sampling is as follows:

- Start with an constant intercept model with $k = 0$ and $\Theta = (1, \ldots, 1)'$.
- Set the initial values of the latent variables $\boldsymbol{\omega}$.
- Draw the intercept (β_0, σ^2) using the update for $(\boldsymbol{\beta}, \sigma^2)$ as given below.
- Start the MCMC sampler and iterate.
    - Draw latent variable $\boldsymbol{\omega}$ given the current model.
    - Update prior precision $\boldsymbol{\lambda}$ on $\boldsymbol{\beta}$ as given below.
    - Update Θ using one of the following moves with equal probability.
        - Add a spline basis function.
        - Delete a spline basis function.
        - Alter a spline basis function.
    - Redraw $(\boldsymbol{\beta}, \sigma^2)$
    - Accept the modifications to Θ and $\boldsymbol{\beta}$ with probability,

      $$Q = \min\left\{1, \frac{|\widehat{V}^*|^{1/2}}{|\widehat{V}|^{1/2}} \exp\left(\frac{a}{a^*}\right)\right\}$$

      where $|\widehat{V}|$ is the determinant of the posterior variance covariance matrix of $\boldsymbol{\beta}$ and is given by $(\Theta'\Theta + D)^{-1}$, the superscript $*$ refer to the parameters of the proposed update model and a is the error term,

$$a = \boldsymbol{\omega}'\boldsymbol{\omega} - \widehat{\boldsymbol{\beta}}'\widehat{V}^{-1}\widehat{\boldsymbol{\beta}}.$$

- Otherwise keep the current model.

The procedures for updating Θ i.e., adding, deleting and modifying a spline base and for updating $(\boldsymbol{\beta}, \sigma^2)$ are given below.

Adding a spline

The steps to add a basis function to the model is as follows:

1. Draw the interaction level of the spline $z_j \sim U(1, \ldots, z_{\text{max}})$.

2. Draw z_j elements $\{d_{j1}, \ldots, d_{jz_j}\}$ from $\{1, \ldots, p\}$ without replacement.

3. For each of the z_j interactions that make up the jth spline: select a data point at random from the data set, say $\boldsymbol{x}_i$ and set the corresponding knot point $\theta_{jl} = x_{id_{jl}}$. Then draw the orientation of the spline from uniform $\{0, 1\}$, where 0 corresponds to $+$ (positive orientation) and 1 to $-$ (negative orientation).

4. Update $(\boldsymbol{\beta}, \sigma^2)$ as given below.

Deleting a spline

Choose one of the k splines at random and remove it from the model and subsequently update the values of $(\boldsymbol{\beta}, \sigma^2)$ as shown below

Modifying a spline

The following is the procedure to modify a basis function to the model,

1. Select at random one of the k splines, say the jth, to modify.

2. Select the lth of the z_j interactions at random and reset the knot point θjl by randomly drawing a data point $\boldsymbol{x}_i$ from the data set and fixing the value of $\theta_{jl} = x_{id_{jl}}$.

3. Update $(\boldsymbol{\beta}, \sigma^2)$ as given below.

Updating the latent variables $\boldsymbol{\omega}$

For the update to $\boldsymbol{\omega}$, we propose to update each ω_i in turn conditional on the rest. That is, we update $\omega_i | \boldsymbol{\omega}_{-i}, \mathbf{Y}, \mathcal{M}$ $(i = 1, \ldots, n)$, where $\boldsymbol{\omega}_{-i}$ indicates the $\boldsymbol{\omega}$ with the ith element removed.

The latent variables ω_i's conditional on the current model parameters $\mathcal{M}$ and the data Y_i does not have an explicit form. Thus we resort to the Metropolis-Hastings procedure with a proposal density $T(\omega_i^* | \omega_i)$ that generates the moves from the current state ω_i to a new state ω_i^*. The proposed updates are then accepted with probabilities,

$$\alpha = \min \left\{ 1, \frac{p(y_i|\omega_i^*) p(\omega_i^* | \boldsymbol{\omega}_{-i}, \Theta) T(\omega_i | \omega_i^*)}{p(y_i|\omega_i) p(\omega_i | \boldsymbol{\omega}_{-i}, \Theta) T(\omega_i^* | \omega_i)} \right\},$$

otherwise the current model is retained.

Finally, the full conditional for ω_i is,

$$p(\omega_i | \boldsymbol{\omega}_{-i}, \mathbf{Y}, \mathcal{M}) \propto \exp \left[\sum_{j=1}^{n} Y_i \omega_i - \sum_{j=1}^{n} \log(1 + \exp(\omega_i)) - \frac{1}{2\sigma^2} (\omega_i - \Theta_i' \boldsymbol{\beta})^2 \right]$$

where Θ_i is the ith row of MARS basis matrix Θ as given in (3.4).

It is convenient to take the proposal distribution $T(\omega_i^* | \omega_i)$ to be a symmetric distribution (eg. Gaussian) with mean equal to the old value ω_i and a pre-specified standard deviation.

Updating $(\boldsymbol{\beta}, \sigma^2)$ conditional on changes to the spline base and latent variables $\boldsymbol{\omega}$

Conditional on the latent variables $\boldsymbol{\omega}$ and the current MARS model, using Bayesian linear model theory we update the spline coefficients and the residual random effects, given the changes to the spline basis using their posterior distribution, so that

$$(\boldsymbol{\beta}, \sigma^2) \quad \sim \quad \mathrm{N}_{n+1}(\boldsymbol{\beta} | \boldsymbol{m}, \sigma^2 \mathbf{V}) \mathrm{IG}(\sigma^2 | \tilde{\gamma_1}, \tilde{\gamma_2}),$$

where $m = V(\Theta^*)'\boldsymbol{\omega}$, $V = [(\Theta^*)'\Theta^* + \boldsymbol{D}]^{-1}$, $\tilde{\gamma_1} = (\gamma_1 + n/2)$, and $\tilde{\gamma_2} = (\gamma_2 + (1/2)(\boldsymbol{\omega}'\boldsymbol{\omega} - m'Vm))$. Here Θ^* now is the $n \times (k+1)$ matrix of outputs from k splines with the intercept and $\boldsymbol{D}$ is the prior precision on $\boldsymbol{\beta}$.

Updating prior precision $\boldsymbol{\lambda}$ conditional on the current model

We draw new values of λ_i using the conditional posterior disribution,

$$\lambda_i \quad \sim \quad \mathrm{Gamma}(\tau_1 + \frac{1}{2}, \tau_2 + \frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2})$$

where k is the number of basis functions and $\boldsymbol{\beta}$ are the regression coefficients.

VITA

Veerabhadran Baladandayuthapani, son of Pushpa Dandapani and V.B. Dandapani, was born in Durgapur, West Benagal, India on November 18th 1976. He received his high school education from St. Francis College, Lucknow, India. He received his Bachelor of Science (Honors) in Mathematics from the Indian Institute of Technology, Kharagpur, India in 1999. That same year he joined Department of Biostatistics, University of Rochester, New York and received a Master of Arts in Statistics in 2000. He continued his studies in statistics at Texas A&M University under the direction of Distinguished Professor Raymond J. Carroll and Professor Bani K. Mallick, and received his Doctor of Philosophy in December 2005. He is employed as a tenure track assistant professor at M.D. Anderson Cancer Center in the Department of Biostatistics & Applied Mathematics. His permanent address is

Department of Applied Mathematics and Biostatistics
1515 Holcombe Blvd, Houston, TX 77030