

Nonparametric Bayesian Models for Supervised Dimension Reduction and Regression

by

Kai Mao

Department of Statistical Science
Duke University

Date: _____

Approved:

Shayan Mukherjee, Supervisor

Feng Liang

Mike West

Merlise Clyde

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

2009

ABSTRACT

Nonparametric Bayesian Models for Supervised Dimension
Reduction and Regression

by

Kai Mao

Department of Statistical Science
Duke University

Date: _____

Approved:

Shayan Mukherjee, Supervisor

Feng Liang

Mike West

Merlise Clyde

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

2009

Copyright © 2009 by Kai Mao
All rights reserved

Abstract

We propose nonparametric Bayesian models for supervised dimension reduction and regression problems. Supervised dimension reduction is a setting where one needs to reduce the dimensionality of the predictors or find the dimension reduction subspace and lose little or no predictive information. Our first method retrieves the dimension reduction subspace in the inverse regression framework by utilizing a dependent Dirichlet process that allows for natural clustering for the data in terms of both the response and predictor variables. Our second method is based on ideas from the gradient learning framework and retrieves the dimension reduction subspace through coherent nonparametric Bayesian kernel models. We also discuss and provide a new rationalization of kernel regression based on nonparametric Bayesian models allowing for direct and formal inference on the uncertain regression functions. Our proposed models apply for high dimensional cases where the number of variables far exceed the sample size, and hold for both the classical setting of Euclidean subspaces and the Riemannian setting where the marginal distribution is concentrated on a manifold. Our Bayesian perspective adds appropriate probabilistic and statistical frameworks that allow for rich inference such as uncertainty estimation which is important for measuring the estimates. Formal probabilistic models with likelihoods and priors are given and efficient posterior sampling can be obtained by Markov chain Monte Carlo methodologies, particularly Gibbs sampling schemes. For the supervised dimension reduction as the posterior draws are linear subspaces which are points on a Grassmann manifold, we do the posterior inference with respect to geodesics on the Grassmannian. The utility of our approaches is illustrated on simulated and real examples.

Contents

Abstract	iv
List of Figures	viii
List of Tables	xi
Acknowledgements	xii
1 Introduction	1
1.1 Dirichlet Processes and Extensions	4
1.1.1 Dirichlet Processes	4
1.1.2 Dependent Dirichlet Processes	8
1.2 Kernel Models	9
1.3 Dimension Reduction	11
1.3.1 The Setup	11
1.3.2 Some Dimension Reduction Techniques	12
1.3.3 Supervised Dimension Reduction	14
1.4 Overview of the Remaining Chapters	17
2 Bayesian Mixture Modeling of Inverse Regression for Supervised Dimension Reduction	19
2.1 Introduction	19
2.2 Bayesian mixtures of inverse regression (BMI)	22
2.2.1 Model specification	22
2.2.2 Inference on the Model Parameters	26
2.2.3 Posterior Inference on the d.r. subspace	31
2.2.4 The $p \gg n$ setting	33

2.2.5	Selecting d	33
2.3	Application to simulated and real data	34
2.3.1	Regression on a nonlinear manifold	35
2.3.2	Classification	39
2.3.3	High-dimensional data: digits	40
2.4	Discussion and Future Work	42
3	Bayesian Gradient Learning Through Kernel Models for Supervised Dimension Reduction	44
3.1	Introduction	44
3.2	Dimension reduction and conditional independence based on gradients	46
3.2.1	Euclidean setting	46
3.2.2	Manifold setting	47
3.2.3	Conditional independence	48
3.3	Bayesian Gradient Learning Through Nonparametric Kernel Models	49
3.3.1	The model	49
3.3.2	Sampling from the posterior	53
3.3.3	Posterior Inference	55
3.3.4	Binary regression	56
3.3.5	Selecting d	58
3.3.6	Modeling comments	58
3.4	Simulated and real data examples	59
3.4.1	Linear regression and dimension reduction	59
3.4.2	Linear regression and graphical models	60
3.4.3	Digits analysis	61

3.4.4	Inference of graphical models for cancer progression	64
3.5	Discussion	65
4	Nonparametric Bayesian Kernel Models for Regression and Classification	67
4.1	Introduction	67
4.2	A Class of Non-Parametric Bayesian Kernel Models	69
4.2.1	Problems with Direct Prior Elicitation	69
4.2.2	Priors and Integral Operators	71
4.2.3	Dirichlet Process Priors	71
4.3	Estimation and Inference	73
4.3.1	Likelihood and Prior Specification for Hyper-Parameters	73
4.3.2	Model Fitting and Prediction via MCMC	75
4.3.3	Binary Regression for Classification	76
4.4	Variable and Feature Selection	77
4.4.1	Kernel Model Extension	77
4.4.2	Overall MCMC	78
4.5	Examples	82
4.5.1	Simulated nonlinear classification – variable selection	82
4.5.2	High-dimensional gene expression data – uncertainty in predictions	83
4.5.3	UCI Data sets	85
4.5.4	Modeling considerations	90
4.6	Summary Comments	92
	Appendices	94
	References	98
	Biography	106

List of Figures

2.1	Swiss Roll data: Illustration.	35
2.2	Swiss Roll data: Accuracy for different methods.	37
2.3	The boxplot for the distances between the posterior samples and their Karcher mean.	37
2.4	Cluster labels for all the 200 samples at the last iteration under an experiment. The samples are ordered in terms of the magnitude of Y . Closer samples (in terms of Y) seem to have similar underlying clustering distributions which change "gradually" with increasing response.	38
2.5	Swiss Roll data: Error v.s. number of d.r directions kept. The minimum one corresponds to $d = 3$, the true value.	39
2.6	Visualization of the embedded <i>Iris</i> data onto a 2 dimension subspace. . . .	40
2.7	Visualization of the embedded <i>Iris</i> data onto a 1 dimensional subspace. . . .	41
2.8	Visualization of the <i>Iris</i> data for different methods. (see also Sugiyama (2007) Figure 6.)	41
2.9	BMI: (a) The posterior mean of the top d.r. direction for 3 versus 8, shown in a 28×28 pixel format. (b) The posterior mean of the top d.r. direction for 5 versus 8, shown in a 28×28 pixel format. Difference between digits is reflected by the red color.	42
3.1	(a) The data matrix with rows corresponding to samples and columns to predictor variables (dimensions); (b) The posterior mean of the gradient outer product matrix; (c) The posterior mean of the top d.r. direction; (d) The trace plot for the inner product of two consecutive draws of the top d.r. direction.	60
3.2	(a) and (b) are the posterior mean and standard deviation for the GOP, respectively; (c) and (d) are the posterior mean and standard deviation for the partial correlation matrix, respectively.	62

3.3	Graphical models inferred from the (a) the gradient outer product matrix and (b) the covariance matrix of the predictor variables. Each node represents a variable and each edge indicates conditional dependence. The distance of the edge is inversely proportional to the amount of dependence, the thickness of the edge is proportional to the certainty of the inference and blue edges are negative while red edges are positive.	62
3.4	(a) The posterior mean of the d.r. direction for 3 versus 8, shown in a 28×28 pixel format. (b) The posterior mean of the d.r. direction for 5 versus 8, shown in a 28×28 pixel format.	63
3.5	The association graph for the progression of prostate cancer from benign to malignant based on the inferred partial correlation. Red edges correspond to positive partial correlations and blue for negative. The width of the edges correspond to the degree of uncertainty, edges we are more sure of are thicker.	65
4.1	(a) Illustration for the first two dimensions of the data. Points from class 1 are red “pluses” and points from class 0 are blue “stars”. (b) The posterior means for ρ : the first two signal dimensions have large values. (c) The posterior prediction probability <i>with</i> variable selection over all points in the first two dimensions, $[-2, 2] \times [2, 2]$. (d) The same as (c) but <i>without</i> variable selection.	84
4.2	The posterior predictive distribution for a test set with the first 10 samples are normal and the remaining tumor. The red stars represent the posterior means and the blue lines are 90% credible intervals. There are 13 cases that are misclassified and 7 more that are very uncertain.	86
4.3	Posterior predictive probabilities for belonging to the benign class for 153 test samples. The blue line segments represent 95% credible intervals and the red star is the posterior mean. The first 88 samples are benign and the remaining 65 are malignant.	87
4.4	Boxplots of the posterior distribution of ρ_1, \dots, ρ_9	88
4.5	(a) The posterior predictive probability of a sample being good on the test samples. The first 125 samples are good and the remaining 75 are bad. Red stars are posterior means and blue lines are 95% credible intervals. (b) A boxplot showing the relevant significance for the explanatory variables. . .	89

4.6 Boxplots for the posterior draws for ρ_1 (upper panel) and the posterior probability of the first benign sample under different hyper-parameter values of a_s . The predictions shown are similar, but the distribution of ρ_1 differs greatly due to outliers under larger a_s values. 91

List of Tables

4.1	The training and test accuracies for different methods on the gene expression data	85
4.2	The training and test accuracies for different methods on the Wisconsin data	86
4.3	The training and test accuracies for different methods on the Ionosphere data	88

Acknowledgements

I am deeply grateful to my supervisor Prof. Sayan Mukherjee for his support and guidance. He has given me invaluable instructions and constantly sparked me with great ideas. He is also a remarkably nice and thoughtful person. It has been a great honor for me to work with him all these years. I would also like to thank Prof. Feng Liang, my first year advisor, for her continually insightful guidance even after she moved to UIUC, and my successful passing of the preliminary test was contributed greatly to her detailed advice.

I warmly thank Prof. Mike West for being on my committee and offering me very helpful comments, for providing me with funding for the research of kernel models, and for recommending me to my internship opportunities. I also owe my gratitude to Prof. Merlise Clyde and Prof. Robert Wolpert who offered outstanding teaching for first year basic statistical courses which have fundamentally shaped my statistical thoughts. My thanks go also to all professors who have taught and instructed me and given me valuable advice.

I sincerely thank my family and friends for their love, understanding and support. My mother and my girlfriend Pengpeng have always been accompanying me especially when I was in my lows. I have met a lot of friends at Duke, and they will certainly be my life-long assets.

Chapter 1

Introduction

Nonparametric Bayesian models refer to probability models on function spaces. Unlike traditional parametric models that utilize a finite number of parameters to represent the unknown hence can suffer from either over-fitting or under-fitting when the number of parameters or model complexity is not appropriately specified, in nonparametric Bayesian models the unknown is represented by an infinite dimensional parameter, that is, a function f , for which suitable prior information can be incorporated to control the model complexity. Such a methodology overcomes the rigid nature of parametric assumptions and leads to highly flexible inference: on one hand with the potentially massively many parameters the model support is wide enough to avoid under-fitting, and on the other hand proper choice of priors controls the model complexity hence mitigates over-fitting.

In principle nonparametric Bayesian models find applications anywhere a function f needs to be learned. In unsupervised framework this could be a density estimation task where the target is a density or distribution function of interest, that is, given samples $\theta_1, \dots, \theta_n$ i.i.d. $\sim f$ to make inference on f . Traditional parametric models such as the finite Gaussian mixture modeling approach for density estimation directly assume a specific form for f through pre-specified Gaussian kernel functions and a finite number of parameters. Nonparametric methods, in contrast, place appropriate priors on f . Popular prior choices in this category include Dirichlet Process (DP) priors (Ferguson, 1973, 1974) and their extension stick-breaking priors (Sethuraman, 1994; Ishwaran and James, 2001), Polya Trees (Lavine, 1992, 1994), Logistic Gaussian Processes (Lenk, 1988), Bernstein Polynomials (Petrone, 1999a,b; Petrone and Wasserman, 2002), etc, among which DP priors are

the most popular and have broad applications in a variety of disciplines.

In supervised scenario where there are predictor and response variables available and prediction is of the central focus, f could naturally be the underlying regression or classification function. In regression problems the main purpose is to infer the relationship between a response variable $Y \in \mathcal{Y} \subset \mathbb{R}$ and predictors or explanatory variables $X \in \mathcal{X} \subset \mathbb{R}^p$. The regression model is typically summarized by $Y = f(X, \varepsilon)$ with f an unknown regression function and ε some noise and the goal is to infer the regression function f which represents the relationship between the response and predictor variables. Traditional parametric models proceed by directly assuming a parametric form for f , for instance a linear form, followed by parameter estimation through procedures such as least squares or maximum likelihood methods. Nonparametric Bayesian models, on the other hand, incorporate suitable prior information on f to ensure that f appropriately reflects the relationship between Y and X and avoids both over-fitting and under-fitting. The underlying idea of nonparametric Bayesian models in this setting resembles much of the (non-probabilistic) regularization methodology (Tikhonov and Arsenin, 1977) that generally chooses f from a candidate functional space by a “minimizing loss function plus penalty” framework popular in machine learning community. Indeed they both share the key idea of ensuring wide support (viewing f as a random function from a functional space v.s. choosing f from a candidate functional space) and controllable complexity (placing priors on f v.s. imposing penalty on f). The primary difference is that the former is probabilistic while the latter is not. In cases where probabilistic inference such as evaluation of estimate uncertainty is needed, nonparametric Bayesian models will be highly preferred. Nonparametric approaches have been extensively applied in supervised settings which have greatly facilitated the development of both Statistics and Machine Learning. Popular methods such as kernel models (Vapnik, 1998; Schölkopf and Smola, 2001), spline models (Wahba, 1990), tree models (Breiman et al., 1984; Friedman, 1991), wavelet models (Donoho and John-

stone, 1994) are now standard. Kernel models, with the celebrated “Support Vector Machine” (Cortes and Vapnik, 1995) as a special case, are extensively applied due to their high predictive accuracy and computational ease.

The primary focus of this thesis is to develop nonparametric Bayesian models in dimension reduction and regression problems. Dimension reduction is an effective methodology to mitigate the “curse of dimensionality” (Bellman, 1961) issue when analyzing high dimensional data sets which have been common nowadays in almost all disciplines. Dimension reduction techniques can help researchers gain insight for the structure of the data and problem, overcome the difficulties caused by high dimension such as over-fitting, facilitate the use of other statistical methods which may only be applicable in small or moderate size problems, and reduce data storage cost. Though many dimension reduction methods exist in literature (Cox and Cox, 2001; Mika et al., 1999; Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Li, 1991, 1992; Wu et al., 2007; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Wu et al., 2007, 2008; Xia et al., 2002; Sugiyama, 2007), most of these methods are not probabilistic hence key question of uncertainty estimation cannot be addressed, calling for the imperative need for probabilistic dimension reduction methodologies. Nonparametric Bayesian approaches have been very successful in many fields by bringing about rich inference, however their development and application on dimension reduction tasks are extremely limited (for a recent example see Tokdar et al. (2008)). This thesis aims to bridge the gap between these areas. We develop nonparametric Bayesian models, with focus on Dirichlet Process models and their extensions, and kernel models, in dimension reduction and regression problems. The dimension reduction studied in this thesis will primarily be in the supervised framework in which the reduction is in the sense of preserving the predictive information. We first give a detailed background introduction of Dirichlet Processes related models, kernel models, and the theory and practice of dimension reduction.

1.1 Dirichlet Processes and Extensions

1.1.1 Dirichlet Processes

In the task of density estimation, a standard finite Gaussian mixture modeling approach proceeds by assuming a parametric form through Gaussian kernel functions for the target distribution or density function that is of interest. In this setting however a major limitation is that one needs to specify the number of mixture component which is generally unknown. Nonparametric Bayesian models overcome this limitation by placing proper functional priors on the target unknown distribution function, so that inference could be done directly on the target distribution function itself, henceforth adding in more flexibility.

Among all functional priors over distribution functions, the Dirichlet process (DP) prior is perhaps the most popular choice. Suppose random variables $\theta_1, \dots, \theta_n$ which could be either samples or parameters follow a distribution function G . Consider placing a prior on G . For a specified distribution G_0 called “base measure” having the same support as G and a positive scale/concentration parameter α_0 , the notation $\text{DP}(\alpha_0, G_0)$ implies that for any measurable partition of the sample space (B_1, B_2, \dots, B_k) , the random vector $(G(B_1), \dots, G(B_k))$ follows a Dirichlet distribution, that is

$$(G(B_1), \dots, G(B_k)) \sim \text{Dirichlet}(\alpha_0 G_0(B_1), \dots, \alpha_0 G_0(B_k))$$

(Ferguson, 1973, 1974). This definition implies that the Dirichlet process is a natural extension to random functions of the traditional Dirichlet distribution on finite dimensional random variables. Here $E(G) = G_0$, so that the base measure serves as the “center” for the random distribution G ; The scale parameter α_0 reflects the concentration of G to G_0 : in fact one has $\text{Var}(G(A)) = G_0(A)(1 - G_0(A))/(\alpha_0 + 1)$ for any measurable set A , so that a bigger α_0 implies that G will have a smaller variance hence tend to be more sim-

ilar to G_0 . The technical definition of DP prior however does not demonstrate what the function with such a prior should look like. The two representations or characterizations of the DP, namely, the stick breaking representation (Sethuraman, 1994) and the Polya-urn representation (Blackwell and MacQueen, 1973), provide insight into important properties of DP.

The stick breaking representation states that a distribution function following a DP prior is an infinite mixture

$$\begin{aligned}
 G &= \sum_{h=1}^{\infty} \pi_h \delta_{\nu_h^*}, & (1.1) \\
 \pi_h &= \beta_h \prod_{l=1}^{h-1} (1 - \beta_l), \quad \beta_h \sim \text{Beta}(1, \alpha_0), \\
 \nu_h^* &\sim G_0
 \end{aligned}$$

where δ_{\cdot} is the delta function, ν_h^* 's are called ‘‘atoms’’ that are simply i.i.d. draws from the base measure and π_h 's are called ‘‘weights’’ that are constructed from a unit ‘‘stick’’ by sequentially sampling from a Beta distribution based on what is left after the previous sampling. This construction demonstrates a specific form for the distribution functions that follow a DP prior, and a salient point is that such functions have a discrete support, the random atoms sampled from the base measure G_0 . As a consequence, a natural clustering machinery will be induced for the random variables $\theta_1, \dots, \theta_n$ that follow G . The scale parameter α_0 controls the magnitude of the β_h , so that a big α_0 tends to induce small β_h hence small π_h in general which tends to make G more similar or concentrated to G_0 .

A second representation, the Polya-urn scheme, focuses on the implication for the DP prior on the random variables following G rather than the functional form for G itself. Suppose $\theta_1, \dots, \theta_n \sim G$ and $G \sim \text{DP}(\alpha_0, G_0)$, then the predictive distribution for a new

element θ_{n+1} given $\theta_1, \dots, \theta_n$ when G is integrated out, is given by

$$P(\theta_{n+1} \in A | \theta_1, \dots, \theta_n) = \frac{1}{\alpha_0 + n} (\alpha_0 G_0(A) + \sum_{i=1}^n \delta_{\theta_i}(A)) \quad (1.2)$$

where A is any measure set and $\delta_{\theta_i}(A) = 1$ if $\theta_i \in A$ and $= 0$ otherwise. This predictive form is highly consistent with the discrete nature of G implied by (1.1) and the clustering machinery, in that θ_{n+1} could equal to one of $\theta_1, \dots, \theta_n$ with probability proportional to 1 and also has a probability proportional to α_0 of having a brand new value sampled from the base measure.

The discrete nature of the DP prior might be undesirable in applications where continuous distribution functions are of interest. In such cases one simply needs to let $\theta_1, \dots, \theta_n$ be individual parameters of a continuous function which could be the sampling distribution function. Specifically one could have

$$\begin{aligned} y_i | \theta_i &\sim g(\cdot | \theta_i) \\ \theta_i | G &\sim G \\ G &\sim \text{DP}(\alpha_0, G_0) \end{aligned} \quad (1.3)$$

where y_i 's are the actually observed samples and $g(\cdot | \theta_i)$ is a continuous distribution function in the sample with parameters θ_i . This is termed ‘‘Dirichlet Process mixture models (DPM)’’. For example, if g is assumed to be Gaussian then a Gaussian mixture model is effectively produced and resembles much of the traditional Gaussian mixture modeling framework, for the extra benefit that one lets the data speak for itself and does not have to impose any rigid assumptions such as pre-specifying the number of mixture components as one would have to do in traditional parametric mixture modeling framework.

Computational advantages are also a major factor for the popularity of the DP pri-

ors. Efficient Gibbs sampling methodologies are developed in literature (Escobar and West, 1995; MacEachern and Müller, 1998; Ishwaran and James, 2001), which are primarily divided into two categories: the conditional approach and the marginal approach. The conditional approach is related to the stick breaking representation for DP and in practice aims to approximate the infinite mixtures in (1.1) by a mixture with sufficiently many mixture components, i.e., to find a large enough H to replace ∞ in (1.1). The resulting sampling scheme is then analogous to that for finite mixture models, that is, for $\theta_1, \dots, \theta_n \sim G = \sum_{h=1}^H \pi_h \delta_{\nu_h^*}$, one introduces latent variables called “labels” z_1, \dots, z_n s.t.

$$z_i = h \Leftrightarrow \theta_i = \nu_h^*$$

meaning the i -th sample or parameter is assigned to the h -th mixture component, and iteratively updates z_i, θ_i, ν_h^* under their full conditional posterior distributions which are easily tractable. The marginal approach, on the other hand, utilizes the Polya-urn scheme in (1.2) which implies a prior on $\theta_i (i = 1 \dots, n)$, s.t.

$$\theta_i | \theta_{-i} = \frac{1}{\alpha_0 + n - 1} \left(\alpha_0 G_0 + \sum_{j=1}^{n-1} \delta_{\theta_j} \right)$$

where θ_{-i} denotes $\{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n\}$. If there is a sampling distribution as in (1.3), the sampling distribution is taken as the likelihood function in θ_i , and by multiplying the likelihood and the prior for θ_i one can obtain the conditional posterior of $\theta_i | (y_1, \dots, y_n, \theta_{-i})$ again as a mixture with a tractable form as long as g and G_0 are conjugate (Escobar and West, 1995).

1.1.2 Dependent Dirichlet Processes

An important premise for DP models is that $\theta_1, \dots, \theta_n$ are *exchangeable*, that is any finite permutation of the indices should result in the same joint distribution. This natural holds by De Finetti's theorem (Diaconis and Freedman, 1980) in the setting $\theta_1, \dots, \theta_n$ i.i.d. $\sim G$ which is a commonly adopted framework. In some scenarios, however, θ_i is dependent on covariates say $x_i \in \mathcal{X}$, so that θ_i in effect follows a distribution function G_i that depends on x_i . For example in density estimation tasks where covariates are available, instead of assuming all the samples θ_i 's from the same common distribution G , it is more desirable to assign an individual covariates-dependent distribution function G_i to each θ_i ; Supervised framework where y_i is a response and x_i is an predictor variable also falls into this category: the distribution of y_i is x_i -dependent. In such settings the θ_i 's are no longer exchangeable.

Dependent Dirichlet Processes (DDP) are important extensions to DP that address the above issue involving covariates. The DDP was first introduced in MacEachern (1999) to generate DP to settings where covariates need to be incorporated when modeling a unknown distribution G . Consider the stick-breaking representation for G in (1.1). Now if G depends on some covariates x and one wants to induce dependence among different such G 's through the dependence among different x 's one could effectively achieve this by allowing the π_h 's, or the ν_h^* 's, or both, in (1.1), to depend on x . For example, one could have

$$G_x = \sum_{h=1}^{\infty} \pi_{hx} \delta_{\nu_h^*}$$

in which now the subscript x is added to G and π_h to show their explicit dependence on x , where the dependence structure among G_x 's is induced via the dependence among the weights π_{hx} through x . There are multiple ways to construct such dependent G_x 's leading to different DDP (Dunson and Park, 2008; Gelfand et al., 2005; Griffin and Steel, 2006; Iorio et al., 2004; Dunson et al., 2008). As will be seen DDP is highly relevant and useful

in our supervised dimension reduction task.

1.2 Kernel Models

Regression models are typically summarized by $Y = f(X) + \varepsilon$ with f an unknown regression function, X, Y predictor and response variables respectively and ε noise. Non-parametric non-probabilistic methods on regression problems, instead of imposing a rigid parametric form on f , generally adopt a regularization framework popular in machine learning. In this framework f is chosen from a candidate functional space with regularization, that is, f is selected in such a way that it minimizes a pre-chosen loss function and a penalty term, so that over-complex or non-smooth functions tend to be penalized to mitigate the problem of over-fitting. An important class of such models is the nonparametric kernel models which have been used extensively for classification and regression problems (Hastie et al., 2001; Schölkopf and Smola, 2001). The most well known example is the celebrated support vector machines (SVMs) (Cortes and Vapnik, 1995) which have been high successful in practice since its inception. The appealing properties of nonparametric kernel models are their flexibility and predictive accuracy, and most importantly, their ability to handle high dimensional data.

Nonparametric kernel models proceed by constructing a so called reproducing kernel Hilbert space (RKHS) \mathcal{H}_k (Wahba, 1990) generated by a positive semi-definite kernel function $k(x, u)$ with $x, u \in \mathcal{X}$, the input space, and selecting the estimate in \mathcal{H}_k . \mathcal{H}_k can be formally characterized as the following provided that the kernel function $k(\cdot, \cdot)$ is a Mercer kernel (Mercer, 1909):

$$\mathcal{H}_k = \left\{ f \mid f(x) = \sum_{j=1}^{\infty} a_j \phi_j(x) \quad \text{s.t.} \quad \sum_{j=1}^{\infty} a_j^2 / \lambda_j < \infty \right\},$$

where $\{\lambda_j\}$ and $\{\phi_j(x)\}$ are the eigenvalues and eigenfunctions of the Mercer kernel function $k(\cdot, \cdot)$. Denote $k_x(\cdot) = k(x, \cdot)$ referred to as the kernel function at knot x , the RKHS \mathcal{H}_k could be roughly described as a Hilbert space containing all the linear combinations of kernel functions at any knot in the input space, and their point-wise limits. Thus

$$\mathcal{H}_k = \left\{ \overline{\sum_{t=1}^T a_t k_{x_t}(\cdot)}, a_t \in \mathbb{R}, x_t \in \mathcal{X}, T \in \mathbb{Z}^+ \right\}$$

The regularization or penalized loss function framework can be stated as

$$\hat{f} = \arg \min_{f \in \mathcal{H}_k} L(f, \text{data}) + \lambda \|f\|_{\mathcal{H}_k}^2 \quad (1.4)$$

where L is some loss function measuring the fit of f to the data — typically a more complex (less smooth) form for f would lead to a smaller loss value; the second term is a penalty that penalizes over-complex f that would over-fit the data: $\|\cdot\|_{\mathcal{H}_k}$ is the RKHS norm defined on the RKHS space measuring the complexity of f , and λ is a tuning parameter specifying the strength of the penalty — a larger λ forces the procedure to choose smoother forms for f while a smaller λ encourages more complex and wiggled ones, hence a proper specified λ balances the trade-off between minimizing the fitting errors and the smoothness.

Although the optimization in (1.4) may be over an infinite dimensional space the optimal solution has the following finite dimensional representation due to the representer theorem (Kimeldorf and Wahba, 1971)

$$\hat{f}(x) = \sum_{i=1}^n w_i k(x, x_i), \quad (1.5)$$

This reduces an infinite dimensional optimization problem to one in n variables, which is very attractive for high-dimensional analysis since the optimization is over $n \ll p$ variables

and independent of the dimension p .

While (1.5) is appealing, one can only obtain a point estimate and hence limited inference without a direct way to evaluate estimate and predictive uncertainty unless one seeks additional computationally heavy methods such as bootstrap, a drawback shared by all non-probabilistic methods. A fully Bayesian approach of kernel methods would provide a natural framework to add in probabilistic interpretations and provide rich inference such as easy evaluation of uncertainty for estimates and prediction through posterior samples.

Nonparametric Bayesian kernel methods, for example Bayesian SVMs, have been proposed (Tipping, 2001; Sollich, 2002) based on applying Bayesian estimation directly to the finite representation from equation (1.5). However, the direct adoption of (1.5) is not a proper statistical model, as the model changes with the sample size and observed covariate values defining the knots, without a coherent argument.

1.3 Dimension Reduction

1.3.1 The Setup

High dimensional data sets where one has many variables are common nowadays in diverse domains such as statistics, engineering, bioinformatics, econometrics, etc. Traditional statistical methods typically break down when dealing with these large data sets due to the infamous manifesto “the curse of dimensionality” (Bellman, 1961), which asserts that with the increase of the dimensionality (the number of variables) the available data become sparse exponentially and render many of the traditional methodologies incapable. For example, in a linear regression problem, over-fitting will occur when one has too many predictor variables and make the inference unreliable.

Dimension reduction is a suitable way to overcome this difficulty. It is the process of

reducing the number of variables in consideration and meanwhile aiming to lose as little information as possible. The foundation of dimension reduction is that, in many cases, though the data are from a possibly very high dimensional space, the problem is intrinsically low dimensional. For instance, in a regression problem, although one may potentially have a huge number of predictor variables, only some of them are in fact relevant. Dimension reduction techniques aim to find such a low dimensional structure in these settings, and can thus help researchers better understand the problem and facilitate the use of other statistical methods.

Mathematically the dimension reduction can be stated as follows: given the samples $x_1, \dots, x_n \in \mathcal{X} \subset \mathbb{R}^p$, produce lower dimensional “features” $\tilde{x}_1, \dots, \tilde{x}_n \in \mathbb{X} \subset \mathbb{R}^d$ with $d < p$ according to certain criteria to ensure of minimization of loss of information in certain sense, so that there exists a map

$$D : \mathbb{R}^p \mapsto \mathbb{R}^d, \quad d < p, \quad D(x_i) = \tilde{x}_i$$

Different dimension reduction approaches have different sense of defining the “minimization of loss of information” as will be seen below. Note that the map $D(\cdot)$ may or may not be explicitly found. If $D(\cdot)$ is not explicitly found then only the lower dimensional features $\tilde{x}_1, \dots, \tilde{x}_n$ are produced and one does not have a general map to reduce the dimensionality of new observations.

1.3.2 Some Dimension Reduction Techniques

Dimension reduction techniques are typically classified into linear methods and nonlinear methods. For linear methods the map $D(x_i) = B'x_i$ where $B = (\beta_1, \dots, \beta_d)$ is a $p \times d$ matrix, so that effectively a d dimensional subspace \mathcal{B} formed by the span of the columns of B is found and the p dimensional sample vectors are projected into this subspace to

achieve dimensionality reduction. Principle component analysis (PCA) is perhaps the most celebrated linear dimension reduction technique. PCA seeks the directions or subspace to maximize the variation of the variables, i.e., PCA finds the directions β_1, \dots, β_d s.t.

$$\begin{aligned} \beta_i &= \operatorname{argmax}_{|\beta|=1} \operatorname{Var}(\beta'X), \quad i = 1, \dots, d \\ \text{s.t. } \operatorname{cov}(\beta_i X, \beta_j X) &= 0, \text{ for } j < i, i > 1 \end{aligned}$$

where $X \in \mathcal{X} \subset \mathbb{R}^p$ is a random sample vector. It turns out these directions are simply the first d eigen-vectors of the covariance matrix of X corresponding to the largest d eigen-values. Extensions such as factor models are closely related to dimension reduction. Factor models state that

$$X - \mu = A\tilde{X} + \epsilon$$

where μ is the mean of $X \in \mathcal{X} \subset \mathbb{R}^p$, $A \in \mathbb{R}^{p \times d}$ and $\tilde{X} \in \mathbb{R}^d$ are named “factor loading” matrix and “factor score” respectively, and ϵ is noise with mean 0 and covariance matrix diagonal. The fundamental assertion underlying factor models is that the covariance structure in the p -dimensional variable X is in fact primarily explained or generated by the latent d -dimensional factor score, and the factor loading matrix serves as the measure of correlation between them, which adds in rich interpretations for the framework. This setup is closely related to dimension reduction in that the factor score \tilde{X} can be effectively viewed as the features that are of lower dimensionality than X . Note that in this setting only reduced features are obtained and a general map $D(\cdot)$ is not available.

For nonlinear methods the map $D(\cdot)$ is nonlinear and often cannot be explicitly represented. Standard examples include kernel principle component analysis (Mika et al., 1999) (kernel PCA) and multi-dimensional scaling (MDS) (Cox and Cox, 2001). In kernel PCA one specifies a bi-variate kernel function $k(x, u)$, ($x, u \in \mathcal{X} \subset \mathbb{R}^p$), for instance, $k(x, u) = \exp(-\frac{|x-u|^2}{2\sigma^2})$ with σ being some bandwidth parameter and a kernel matrix $K \in \mathbb{R}^{n \times n}$ s.t.

the (i, j) -th entry is $K_{ij} = k(x_i, x_j)$ where x_i, x_j represent the sample vectors. Then the features are produced in terms of the eigen-vectors of K , i.e., if η_1, \dots, η_d are the d eigen-vectors of K corresponding to the largest d eigen-values with $\eta_j = (\eta_j^{(1)}, \dots, \eta_j^{(n)})'$, then the d dimensional feature vector $\tilde{x}_i = (\eta_1^{(i)}, \dots, \eta_d^{(i)})$, $i = 1, \dots, n$. Clearly the relationship between x_i and \tilde{x}_i is nonlinear and the map $D(\cdot)$ cannot be explicitly given in the sense that for a new observation vector $x \in \mathcal{X}$ one cannot obtain the corresponding feature vector without re-calculating a new kernel matrix and doing an eigen-decomposition on it. Most of the nonlinear dimension reduction approaches in literature are of this flavor, that is they obtain the feature vectors as the eigen-vectors of certain matrix, for instance, multi-dimensional scaling and the recently very popular so-called manifold learning methods such as ISOMAP (Tenenbaum et al., 2000), locally linear embedding (LLE) (Roweis and Saul, 2000), Laplacian eigenmap (Belkin and Niyogi, 2003), etc. These manifold learning methods assert that the high dimensional data in effect lie on an intrinsically low dimensional manifold (not necessarily a subspace) and aim to “learn” this manifold by the local information in the ambient space.

1.3.3 Supervised Dimension Reduction

The methods aforementioned are all unsupervised in that there is no response variable. In settings where one is confronted with both response variables and high dimensional predictor variables one may need to reduce the dimensionality of the predictors while keep the predictive information. Supervised dimension reduction (SDR) formulates the problem as finding a low-dimensional subspace (or manifold) that contains predictive information of the response variable and seeks to replace the original predictors with projections onto this low dimensional subspace. This low dimensional subspace is often called the dimension reduction (d.r.) space.

The underlying model in supervised dimension reduction is given p -dimensional covariates X and a response Y the following holds

$$Y = g(b'_1 X, \dots, b'_d X, \varepsilon) \quad (1.6)$$

where the column vectors of $B = (b_1, \dots, b_d)$ are named the d.r. directions and ε is noise independent with X . In this framework all the predictive information is contained in the d.r. space \mathcal{B} which is the span of the columns of B , since $Y \perp\!\!\!\perp X \mid P_{\mathcal{B}}X$, where $\perp\!\!\!\perp$ denotes “independent with” and $P_{\mathcal{B}}$ denotes the orthogonal projection operator onto the subspace \mathcal{B} . Such \mathcal{B} may not be unique and one wants to find the so-called central subspace (Cook, 1996) $\mathcal{S}_{Y|X}$ defined as the intersection of all subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ having the property that $Y \perp\!\!\!\perp X \mid P_{\mathcal{S}}X$, so that the central subspace is effectively the “minimal” d.r. subspace. In our work we do not distinguish the term central subspace and d.r. subspace since we always aim for the minimal d.r. subspace.

The d.r. subspace \mathcal{B} is on a so-called Grassmann manifold denoted as $\mathcal{G}_{(d,p)}$ which is defined as the set of all the d dimensional linear subspaces of \mathbb{R}^p . Existing non-Bayesian dimension reduction approaches seek to find a “best” d.r. subspace, which is effectively an optimal point on $\mathcal{G}_{(d,p)}$. However Bayesian probabilistic models are of great interest here since rich inference can be made through posterior distributions on the whole $\mathcal{G}_{(d,p)}$ hence significantly enlarge our scope.

Approaches to inference of the d.r. space in literature can be divided into three categories. The first is called “forward regression”, where the conditional probability $Y \mid X$ is directly modeled, or equivalently in (1.6) the function g is directly modeled. A classic example of this approach is Projection Pursuit Regression (PPR) (Friedman and Stuetzle, 1981), where one assumes an additive form for g in (1.6), namely $Y = \sum_{m=1}^d g_m(b'_m X) + \varepsilon$ and applies a fitting strategy which iteratively performs the task of optimizing over g_m given

b_m by any suitable smoothers such as spline models or kernel models, and optimizing over b_m give g_m which involves a Newton search. A modern Bayesian example is proposed in Tokdar et al. (2008) where a variant of logistic Gaussian processes is utilized to model $Y | P_{\mathcal{B}}X$ or the function g in (1.6). The major limitation of this type of methods is that it is difficult to obtain an accurate estimate for the unknown distribution $Y | P_{\mathcal{B}}X$ or the function g .

The second line of methods is named “gradient learning” approach. Consider a slightly restricted framework in which the reduction is captured in the regression (mean) function: denote the regression function $f(x) = E(Y|X = x)$ and

$$Y = f(X) + \varepsilon = \tilde{g}(b'_1X, \dots, b'_dX) + \varepsilon$$

by additive error assumption. The observation that the gradient of the regression function, $\nabla f \in \mathbb{R}^p$, lies in the d.r. space motivates this approach. A variety of methods exist in this category (Xia et al., 2002; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Wu et al., 2007). In all these methods a central quantity termed “Gradient Outer Product” (GOP) matrix defined as $E(\nabla f \nabla f')$ is estimated and the first d eigen-vectors corresponding to the largest d eigen-values are taken as the basis for the d.r. space. They differ in how inference is done for the gradient or GOP. In Xia et al. (2002) an efficient method named “MAVE” is provided which estimates the gradient ∇f by local polynomial fitting, however it is not directly applicable when the number of predictor or explanatory variables is larger than the sample size due to over-fitting and numerical instability. In Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Wu et al. (2007) kernel methods are utilized to learn the gradient and overcome the over-fitting problems in the $p > n$ scenario by adding a regularization term in the gradient estimate. The utilization of kernels introduces nonlinear features to the problem hence makes these methods more flexible. The main drawback,

however, is their lack of probabilistic interpretations.

The third category is the “inverse regression” type of approaches, where the conditional distribution $X | Y$ is the focus. The classic examples are the celebrated sliced inverse regression (SIR) (Li, 1991) for the continuous response setting and reduced rank linear discriminant analysis (LDA) for the discrete response setting, in which an important idea that the conditional distribution of the predictor given the response can provide useful information in the reduction of the dimensions was introduced. SIR estimates the inverse regression curve $E(X | Y)$ to infer about the d.r. space. SIR and LDA are not probabilistic. These methods are extended to more general probabilistic settings in Principal Fitted Component (PFC) models (Cook, 2007). A major drawback of PFC as well as SIR and LDA is that they suffer from the multi-modality issue that the d.r. space is degenerate when the regression function is symmetric along certain directions in which case important directions might be lost.

1.4 Overview of the Remaining Chapters

In this thesis we develop nonparametric Bayesian models on supervised dimension reduction and regression problems and demonstrate the rich inference brought by these flexible models.

In chapter 2 we shall develop a Bayesian framework for supervised dimension reduction using a flexible nonparametric Bayesian mixture modeling approach that extends the model-based approach of Cook (2007). Our method retrieves the dimension reduction subspace by utilizing a dependent Dirichlet process that allows for natural clustering for the data in terms of both the response and predictor variables. Formal probabilistic models with likelihoods and priors are given and efficient posterior sampling of the d.r. subspace can be obtained by a Gibbs sampler. As the posterior draws are linear subspaces which are

points on a Grassmann manifold, we output the posterior mean d.r. subspace with respect to geodesics on the Grassmannian. The salient point is that our method applies to data generated from distributions where the support of the predictive subspace is not a linear subspace of the predictors but is instead a nonlinear manifold. The projection is still linear but it will contain the nonlinear manifold that is relevant to prediction. The Bayesian formulation of the inverse regression framework has a natural model-based underpinning based on distribution theory.

Chapter 3 will formulate a coherent Bayesian nonparametric model that is based on ideas from the gradient learning framework and adds appropriate probabilistic and statistical frameworks that allow for uncertainty estimation which is important for measuring the estimates. The proposed model holds for both the classical setting of Euclidean subspaces and the Riemannian setting where the marginal distribution is concentrated on a manifold. The method is especially relevant for the high-dimensional setting where the number of predictor variables far exceed the number of observations. A Markov chain Monte Carlo procedure for inference of model parameters is provided. We will illustrate how our Bayesian model allows for formal inference of uncertainty in dimension reduction as well as inference the uncertainty of conditional dependencies in graphical models.

In Chapter 4 a fully Bayesian framework and theory for kernel regression and classification will be provided. We specify priors on the entire RKHS and induce a class of functions that span the RKHS, providing an equivalence between the nonparametric Bayesian models and kernel models used in the penalized loss framework. This implies a Bayesian representer theorem that results in the finite representation in equation (1.5) derived from a Bayesian formulation, and that is coherent across samples and sample sizes. This formal model then easily and coherently addresses problems of inference on hyper-parameters, variable selection, and ancillary issues such as unlabeled data (in semi-supervised learning).

Chapter 2

Bayesian Mixture Modeling of Inverse Regression for Supervised Dimension Reduction

2.1 Introduction

As stated in Section 1.3.3, supervised dimension reduction (SDR) can be formulated as finding a low-dimensional subspace or manifold that contains all the predictive information of the response variable. This low-dimensional subspace is often called the dimension reduction (d.r.) subspace. Projections onto the d.r. space can be used to replace the original predictors, without affecting the prediction. This is a counterpart of unsupervised dimension reduction such as principal components analysis which does not take into account the response variable. The underlying model in supervised dimension reduction is given in (1.6).

A variety of methods for SDR have been proposed in literature that can be divided into three categories: methods based on forward regression that directly model the conditional probability $Y | X$; methods based on learning gradients of the regression function; and methods based on inverse regression that model the conditional distribution $X | Y$. See Section 1.3.3 for details. In this chapter we focus on the inverse regression category.

The idea that the conditional distribution of the predictors given the response can provide useful information in the reduction of the dimensions was introduced in sliced inverse regression (SIR) (Li, 1991) for the regression setting and reduced rank linear discriminant analysis for the classification setting. SIR proposes the semiparametric model in (1.6) and

claims that the centered conditional expectation $E(X | Y = y) - E(X)$, called the inverse regression curve, is contained in the (transformed) d.r. space spanned by the columns of B . SIR is not a model based approach in the sense that a sampling or distributional model is not specified for $X | Y$. The idea of specifying a model for $X | Y$ is developed in principal fitted component (PFC) models (Cook, 2007). Specifically, the PFC model assumes the following multivariate form for the inverse regression

$$X_y = \mu + A\nu_y + \varepsilon \quad (2.1)$$

where $X_y \equiv X | Y = y$; $\mu \in \mathbb{R}^p$ is an intercept; $\varepsilon \sim N(0, \Delta)$ with $\Delta \in \mathbb{R}^{p \times p}$ is a random error term; $A \in \mathbb{R}^{p \times d}$ and $\nu_y \in \mathbb{R}^d$ imply that the mean of the (centered) X_y lie in a subspace spanned by the columns of A with ν_y the coordinate (similar to a factor model setting with A the factor loading matrix and ν_y the factor score). Under this model formulation it is important that ν_y is modeled otherwise the above model is an adaptation of principal components regression, see sections 2.2.1 and 2.2.1 for the models used here. In this framework it can be shown $B = \Delta^{-1}A$ (Cook, 2007), so that the columns of $\Delta^{-1}A$ spans the d.r. space.

SIR and PFC both suffer from the problem that the d.r. space is degenerate when the regression function is symmetric along certain directions of X , in this case important directions might be lost. The primary reason for this is that X_y for certain values of y may not be unimodal: there may be two clusters or components in the conditional distribution $X | Y = y$. An additional drawback of SIR is that the slicing procedure on the response variable is rigid and not based on a distributional model. Intuitively, the slicing approach should allow for borrowing information across the response variable. Data points with similar responses tend to have similar conditional distributions yet because of the rigid nature of the slicing procedure these data points may belong to different bins and are treated

independently.

Sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) utilizes both the first and second moment of the distribution $X|Y$ in capturing the d.r. space however can also be problematic when moments are degenerate as relevant directions are often lost; Principal hessian directions (PHD) (Li, 1992) focuses on the hessian matrix of the regression function $f(x) = E(Y|X = x)$ and locates the directions along which $f(x)$ shows large curvature, however PHD is constrained due to the requirement of strong distributional assumptions such as normality of the predictor variables to accurately estimate the Hessian matrix. Another problem with PHD is that directions that are linearly correlated to the output variables are lost.

To address the multimodality problem a possibility is to develop a mixture model, that is, to assume a normal mixture model rather than a simple normal model for X_y . This is the approach taken in mixture discriminant analysis (MDA) (Hastie and Tibshirani, 1996b) which utilizes in the classification setting a finite Gaussian mixture model for each class. However MDA can only be applied when the response is discrete rather than continuous, and the pre-specification of the (generally unknown) number of mixture components is an issue.

In this chapter we develop a Bayesian methodology we call Bayesian mixtures of inverse regression (BMI) that extends the model-based approach of Cook (2007). A semi-parametric model will be stated. A salient point is that it applies to data generated from distributions where the support of the predictive subspace is not a linear subspace of the predictors but is instead a nonlinear manifold. The projection is still linear but it will contain the nonlinear manifold that is relevant to prediction. The Bayesian formulation of the inverse regression framework has a natural model-based underpinning based on distribution theory. A further important point of great interest is that the d.r. subspace is on a so-called Grassmann manifold denoted as $\mathcal{G}_{(d,p)}$ which is defined as the set of all the d

dimensional linear subspaces of \mathbb{R}^p , and our model allows for rich inference such as uncertainty evaluation by drawing posterior samples (subspaces) from this manifold rather than merely obtaining an optimal point from this manifold as by other SDR methods.

2.2 Bayesian mixtures of inverse regression (BMI)

2.2.1 Model specification

We propose a semiparametric mixture model that generalizes the PFC model (2.1):

$$X \mid (Y = y, \mu_{yx}, \Delta) \sim N(\mu_{yx}, \Delta) \quad (2.2)$$

$$\mu_{yx} = \mu + A\nu_{yx} \quad (2.3)$$

$$\nu_{yx} \sim G_y \quad (2.4)$$

where $\mu \in \mathbb{R}^p$, $\Delta \in \mathbb{R}^{p \times p}$, $A \in \mathbb{R}^{p \times d}$ have the same interpretations as in (2.1); $\nu_{yx} \in \mathbb{R}^d$ is analogous to ν_y in (2.1) except it depends on y and the marginal distribution of X , and it follows a distribution G_y that depends on y . Note that (2.1) can be recovered by assuming $G_y = \delta_{\nu_y}$ which is a point mass at ν_y , and in this case $\nu_{yx} \equiv \nu_y$.

However by considering G_y as a random process hence specifying flexible nonparametric models for $X \mid Y$ we can greatly generalize (2.1). For example a Dirichlet process prior (DP) (Ferguson, 1973, 1974; Sethuraman, 1994) on G_y leads to a mixture model for $X \mid Y$ due to its discrete property and alleviates the need to prespecify the number of mixture components for $X \mid Y$. In the setting of a continuous response the dependent Dirichlet process (DDP) (MacEachern, 1999; Dunson and Park, 2008) can be used to allow dependence between G_y 's.

Proposition 1. For this model the d.r. space is the span of $B = \Delta^{-1}A$, i.e.,

$$Y | X = Y | (\Delta^{-1}A)'X.$$

Proof. Assume in the following A and Δ are given. Assume in (2.3) $\mu = 0$ w.o.l.g. so that $\mu_{yx} = A\nu_{yx}$. Let $p(y|x)$ be the distribution of Y given X . Then

$$\begin{aligned} p(y | x) &= \frac{p(x | y)p(y)}{p(x)} = \frac{p(y)}{p(x)} \int N(x; \mu_{yx}, \Delta) d\pi(\mu_{yx}) \\ &\propto p(y) \int \exp\left(-\frac{1}{2}(x - \mu_{yx})' \Delta^{-1}(x - \mu_{yx})\right) d\pi(\mu_{yx}) \\ &\propto p(y) \exp\left(-\frac{1}{2}(x - P_A x)' \Delta^{-1}(x - P_A x)\right) \int \exp\left(-\frac{1}{2}(P_A x - \mu_{yx})' \Delta^{-1}(P_A x - \mu_{yx})\right) \\ &\quad \exp\left(-\frac{1}{2}(P_A x - \mu_{yx})' \Delta^{-1}(x - P_A x)\right) d\pi(\mu_{yx}) \end{aligned}$$

where $P_A x$ denotes the projection of x onto the column space of A under the Δ^{-1} inner product, i.e.,

$$P_A x = A(A' \Delta^{-1} A)^{-1} A' \Delta^{-1} x.$$

Since μ_{yx} is in the column space of A , the cross term $(P_A x - \mu_{yx})' \Delta^{-1}(x - P_A x) = 0$, which could also be derived by checking that $\mu_{yx} = P_A \mu_{yx}$ and $P_A' \Delta^{-1}(x - P_A x) = 0$.

So that

$$p(y | x) \propto p(y) \int \exp\left(-\frac{1}{2}(P_A x - \mu_{yx})' \Delta^{-1}(P_A x - \mu_{yx})\right) d\pi(\mu_{yx})$$

thus x comes into play only through $A' \Delta^{-1} x$. □

Given data $\{(x_i, y_i)\}_{i=1}^n$ the following sampling distribution is specified from (2.2) -

(2.4)

$$\begin{aligned}x_i \mid (y_i, \mu, \nu_i, A, \Delta) &\sim N(\mu + A\nu_i, \Delta) \\ \nu_i &\sim G_{y_i}\end{aligned}$$

where $\nu_i := \nu_{y_i x_i}$ and the likelihood

$$\begin{aligned}p(x_1, \dots, x_n \mid y_1, \dots, y_n, \mu, A, \Delta, \nu_1, \dots, \nu_n) &\propto \\ \det(\Delta^{-1})^{\frac{n}{2}} \exp \left[-\frac{1}{2} \sum_{i=1}^n (x_i - \mu - A\nu_i)' \Delta^{-1} (x_i - \mu - A\nu_i) \right] &\quad (2.5)\end{aligned}$$

where $\nu = (\nu_1, \dots, \nu_n)$. To fully specify the model we need to specify the distributions G_{y_i} . The categorical response case is specified in subsection 2.2.1 and the continuous response case is specified in subsection 2.2.1.

Categorical response

When the response is categorical, $y = \{1, \dots, C\}$, we can specify the following model for ν_i

$$\nu_i \mid (y_i = c) \sim G_c \quad \text{for } c = 1, \dots, C, \quad (2.6)$$

where each G_c is an unknown distribution independent with each other. It is natural to use a Dirichlet process as a prior for each G_c

$$G_c \sim \text{DP}(\alpha_0, G_0) \quad (2.7)$$

with α_0 is a concentration parameter and G_0 the base measure. The discrete nature of the DP will ensure a mixture representation for G_c and induce a mixture of normal distributions for $X \mid Y$. This allows for multiple clusters in each class.

Continuous response

In the case of a continuous response variable it is natural to expect G_{y_1} and G_{y_2} to be dependent if y_1 is close to y_2 , that is, we would like to borrow information across the response variables. A natural way of doing this is to use a dependent Dirichlet Process (DDP) prior. The DDP was first introduced in MacEachern (1999) to generate DP to settings where co-variables need to be incorporated when modeling an unknown distribution G . Consider the stick-breaking representation (Sethuraman, 1994) for $G \sim \text{DP}(\alpha_0, G_0)$ in (1.1). Now if one wants G to depend on some variables y and induces dependence among different such G 's through the dependence among y 's one could allow the π_h 's, or the ν_h^* 's, or both, to depend on y . For example, one could have

$$G_y = \sum_{h=1}^{\infty} \pi_{hy} \delta_{\nu_h^*}$$

in which now the subscript y is added to G and π_h to show their explicit dependence on y , where the dependence structure among G_y 's is induced via the dependence among the weights π_{hy} through y . There are multiple ways to construct such dependent G_y 's leading to different DDP (Dunson and Park, 2008; Gelfand et al., 2005; Griffin and Steel, 2006; Iorio et al., 2004; Dunson et al., 2008). Here we utilize the kernel stick breaking process (Dunson and Park, 2008) due to its nice properties and computational efficiency. The kernel stick breaking process constructs G_y in such as way that

$$G_y = \sum_{h=1}^{\infty} U(y; V_h, L_h) \prod_{\ell < h} (1 - U(y; V_\ell, L_\ell)) \delta_{\nu_h^*} \quad (2.8)$$

$$U(y; V_h, L_h) = V_h K(y, L_h) \quad (2.9)$$

where L_h is a random location in the domain of y , $V_h \sim \text{Be}(v_a, v_b)$ a prior is a probability weight, ν_h^* is an atom, and $K(y, L_h)$ is a kernel function that measures the similarity

between y and L_h . Examples of K are

$$K(y, L_h) = 1_{|y-L_h|<\phi} \quad \text{or} \quad K(y, L_h) = \exp(-\phi|y - L_h|^2). \quad (2.10)$$

Dependence on the weights $U(y; V_h, L_h)$ in (2.8) will result in dependence between G_{y_1} and G_{y_2} when y_1 and y_2 are close.

2.2.2 Inference on the Model Parameters

Given data $\{(x_i, y_i)\}_{i=1}^n$ we would like to infer the model parameters $A, \Delta, \nu \equiv (\nu_1, \dots, \nu_n)$. From A and Δ we can compute the d.r. which is the span of $B = \Delta^{-1}A$. The inference will be based on Markov chain Monte Carlo (MCMC) samples from the posterior distribution given the likelihood function in (2.5) and suitable prior specifications. The inference procedure is a Gibbs sampling scheme which can be broken into four sampling steps: sampling μ , sampling A , sampling Δ^{-1} , and sampling ν . The fourth step will differ based on whether the response variable is continuous or categorical.

Sampling μ

A noninformative prior on the intercept parameter μ , i.e., $\mu \propto 1$, combined with the likelihood function (2.5), leads to normal full conditional posterior distribution

$$\mu \mid (\text{data}, A, \nu) \sim N \left(\frac{1}{n} \sum_{i=1}^n (x_i - A\nu_i), \frac{1}{n} \Delta \right)$$

Sampling A

The matrix $A \in \mathbb{R}^{p \times d}$ represents the transformed d.r. space and the likelihood (2.5) implies a normal form in A . We will use the Bayesian factor modeling framework developed in Lopes and West (2004) to model A . In this framework A is viewed as a factor loading

matrix. The key idea in (Lopes and West, 2004) is to impose special structure on A to ensure identifiability

$$A = \begin{pmatrix} a_{11} & 0 & 0 \\ \vdots & \ddots & 0 \\ a_{d1} & \dots & a_{dd} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pd} \end{pmatrix} \quad (2.11)$$

We specify normal and independent priors for the elements of A

$$a_{\ell j} \sim N(0, \phi_a^{-1}), \ell \geq j, \ell = 1, \dots, p$$

the hyper-parameter ϕ_a is specified to take a small value to reflect the vagueness of the prior information.

Conjugacy of the likelihood and the prior leads to a normal conditional posterior for each row of A which we will specify. We first fix notation: the ℓ -th row of A is a_ℓ ; the ℓ -th column of the identity matrix is $I_\ell \in \mathbb{R}^p$; $A_{-\ell} \in \mathbb{R}^{(p-1) \times p}$ is the matrix A with the ℓ -th row removed; $I_{-\ell} \in \mathbb{R}^{p \times (p-1)}$ is the identity matrix with the ℓ -th column removed and

$$\begin{aligned} x_{i/\ell} &= x_i - \mu - I_{-\ell} A_{-\ell} \nu_i, \\ \tilde{\nu}_{\ell,i} &= \begin{cases} \nu_i \equiv (\nu_{i1}, \dots, \nu_{id})', & \ell = d+1, \dots, p \\ (\nu_{i1}, \dots, \nu_{i\ell})', & \ell = 1, \dots, d \end{cases} \end{aligned}$$

The conditional for the ℓ -th row of A is calculated to be

$$\begin{aligned}
a_\ell \mid (\text{data}, A_{-\ell}, \Delta, \nu, \mu) &\sim N(\mu_\ell^{(a)}, \Sigma_\ell^{(a)}) \\
\Sigma_\ell^{(a)} &= [(I_\ell' \Delta^{-1} I_\ell) \sum_i \tilde{\nu}_{\ell,i} \tilde{\nu}'_{\ell,i} + \phi_a \mathbf{I}_{d_\ell^*}]^{-1} \\
\mu_\ell^{(a)} &= \Sigma_\ell^{(a)} \left(\sum_i \tilde{\nu}_{\ell,i} x'_{i/\ell} \right) \Delta^{-1} I_\ell
\end{aligned}$$

where $\mathbf{I}_{d_\ell^*}$ is the $d_\ell^* \times d_\ell^*$ identity matrix with $d_\ell^* = \min(d, \ell)$.

Sampling Δ

A natural choice for a prior for Δ^{-1} is a Wishart distribution $W(df, p, V_D)$ with df degrees of freedom, and scale matrix V_D . This results in the following conditional distribution

$$\begin{aligned}
\Delta^{-1} \mid (\text{data}, A, \nu, \mu) &\propto \det(\Delta^{-1})^{\frac{df-p-1+n}{2}} \\
&\exp \left\{ -\frac{1}{2} \text{Trace} \left((V_D^{-1} + \sum_{i=1}^n (x_i - \mu - A\nu_i)(x_i - \mu - A\nu_i)') \Delta^{-1} \right) \right\}
\end{aligned}$$

Sampling ν for categorical responses

Inference for DP mixture models has been extensively developed in the literature (Escobar and West, 1995; MacEachern and Müller, 1998). We utilize the sampling scheme in Escobar and West (1995) which adopts a marginal approach in sampling from the DP priors. Marginalizing in (2.6) the unknown distribution G_c leads to the poly-urn representation of the prior for ν_i

$$\nu_i \mid (y_i = c, \nu_{-i}) \propto \sum_{j \neq i, y_j = c} \delta_{\nu_j} + \alpha_0 G_0(\nu_i),$$

where $\nu_{-i} = \{\nu_1, \dots, \nu_{i-1}, \nu_{i+1}, \dots, \nu_n\}$, G_0 is the base distribution and α_0 is the base concentration parameter. The fact that ν_i should be constrained to have unit variance to

ensure identifiability implies that a natural choice of G_0 is $N(0, \mathbf{I}_d)$. Combining with the likelihood (2.5) the full conditional for ν_i is

$$\nu_i \mid (\text{data}, y_i = c, \nu_{-i}, A, \Delta, \mu) \propto \sum_{j \neq i, y_j = c} q_{i,j} \delta_{\nu_j} + q_{i,0} G_i(\nu_i)$$

where

$$\begin{aligned} G_i(\nu_i) &\sim N(V_\nu A' \Delta^{-1} (x_i - \mu), V_\nu) \\ q_{i,j} &\propto \exp \left\{ -\frac{1}{2} (x_i - \mu - A\nu_j)' \Delta^{-1} (x_i - \mu - A\nu_j) \right\} \\ q_{i,0} &\propto \alpha_0 V_\nu^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x_i - \mu)' (\Delta^{-1} - \Delta^{-1} A V_\nu A' \Delta^{-1}) (x_i - \mu) \right\} \end{aligned}$$

where $V_\nu = (A' \Delta^{-1} A + \mathbf{I}_d)^{-1}$ and by \propto we mean that $\sum_{j: y_j = y_i} q_{i,j} + q_{i,0} = 1$

Sampling ν for continuous responses

We follow the sampling scheme for the kernel stick-breaking process developed in Dunson and Park (2008). Inference for the DDP is based on a truncation of (2.8)

$$G_y = \sum_{h=1}^H U(y; V_h, L_h) \prod_{l < h} (1 - U(y; V_l, L_l)) \delta_{\nu_h^*}$$

where H some pre-specified value large integer and $U(y; V_h, L_h) = V_h K(y, L_h) = V_h \exp(-\phi |y - L_h|^2)$ for $h = 1, \dots, H-1$ and $U(y; V_H, L_H) = 1$ to ensure that $\sum_{h=1}^H U(y; V_h, L_h) \prod_{l < h} (1 - U(y; V_l, L_l)) = 1$. We denote by K_i the cluster label for sample i , that is, $K_i = h$ means that sample i is assigned to cluster h . To facilitate sampling V_h we introduce latent variables $Q_{ih} \sim \text{Ber}(V_h)$ and $R_{ih} \sim \text{Ber}(K(y_i, L_h))$ for $i = 1, \dots, n$ and $h = 1, \dots, K_i$.

The following iterative procedure provides samples of ν_i

1. Sample the cluster membership K_i

$$K_i = h \mid \text{data}, A, \Delta, \mu, \nu_h^*, V_h, L_h \propto \exp \left\{ -\frac{1}{2}(x_i - \mu - A\nu_h^*)' \Delta^{-1} (x_i - \mu - A\nu_h^*) \right\} \\ \times U(y; V_h, L_h) \prod_{\ell < h} (1 - U(y; V_\ell, L_\ell)) \text{ for } i = 1, \dots, H;$$

This is a multinomial distribution. If the sampled index is h^* then set $\nu_i = \nu_{h^*}^*$.

2. Sample the atoms ν_h^* with prior $\nu_h^* \sim N(0, \mathbf{I}_d)$

$$\nu_h^* \mid \text{data}, A, \Delta, \mu, \sim N \left((n_h Q' \Delta^{-1} A + \mathbf{I}_d)^{-1} A' \Delta^{-1} \sum_{i \in C_h} (x_i - \mu), (n_h A' \Delta^{-1} A + \mathbf{I}_d)^{-1} \right),$$

where C_h is the index for the h -th cluster and n_h is the the cardinality of C_h .

3. Sample V_h with prior $V_h \sim N(v_a, v_b)$

$$V_h \mid \text{data}, Q_{ih}, K_i \sim \text{Be} \left(v_a + \sum_{i: K_i \geq h} Q_{ih}, v_b + \sum_{i: K_i \geq h} (1 - Q_{ih}) \right), \text{ for } h < H \text{ and } V_H \equiv 1.$$

4. Sample the latent variables Q_{ih}, R_{ih} with prior $Q_{ih} \sim \text{Ber}(V_h)$ and $R_{ih} \sim \text{Ber}(K(y_i, L_h))$

$$(Q_{ih} = 1, R_{ih} = 0) \mid V_h, L_h, K_i \sim \frac{V_h(1 - K(y_i, L_h))}{1 - V_h K(y_i, L_h)} \\ (Q_{ih} = 0, R_{ih} = 1) \mid V_h, L_h, K_i \sim \frac{(1 - V_h)K(y_i, L_h)}{1 - V_h K(y_i, L_h)} \\ (Q_{ih} = 0, R_{ih} = 0) \mid V_h, L_h, K_i \sim \frac{(1 - V_h)(1 - K(y_i, L_h))}{1 - V_h K(y_i, L_h)}$$

for $h < K_i$ and $Q_{ih} = R_{ih} = 1$ for $h = K_i$.

5. Sample the locations L_h with non-informative prior $L_h \propto 1$

A Metropolis-Hastings step is taken with the proposal distribution

$$L_h^* \sim \text{Unif}\left(\min_{i \in \{R_{ih}=1\}}(y_i), \max_{i \in \{R_{ih}=1\}}(y_i)\right)$$

In case $\{i : R_{ih} = 1\} = \emptyset$ we sample L_h^* from the prior. The acceptance ratio is calculated to be

$$\prod_{K_i \geq h} \frac{K(y_i, L_h^*)^{R_{ih}} (1 - K(y_i, L_h^*))^{1-R_{ih}}}{K(y_i, L_h)^{R_{ih}} (1 - K(y_i, L_h))^{1-R_{ih}}}$$

6. The kernel precision parameter ϕ in $K(y, L_h) = \exp(-\phi|y - L_h|^2)$ can be pre-specified or sampled. The sampling scheme is as follows:

Let $\tilde{\phi} = \log(\phi)$. By placing a normal prior on $\tilde{\phi}$, namely, $\tilde{\phi} \sim N(\mu_{\tilde{\phi}}, \sigma_{\tilde{\phi}}^2)$ which implies a log-normal prior on ϕ , and a proposal distribution a random walk $\tilde{\phi}^* \sim N(\tilde{\phi}, \sigma_{\text{prop}}^2)$, with $\mu_{\tilde{\phi}}, \sigma_{\tilde{\phi}}^2, \sigma_{\text{prop}}^2$ all pre-specified hyper-parameters, one has the acceptance ratio:

$$\prod_{K_i \geq h} \frac{K^*(y_i, L_h)^{R_{ih}} (1 - K^*(y_i, L_h))^{1-R_{ih}}}{K(y_i, L_h)^{R_{ih}} (1 - K(y_i, L_h))^{1-R_{ih}}}$$

where $K^*(y_i, L_h) = \exp(-\phi^*|y - L_h|^2)$ with the proposed $\phi^* = \exp(\tilde{\phi}^*)$.

2.2.3 Posterior Inference on the d.r. subspace

Given posterior samples of the parameters A and Δ^{-1} and the formula $B = \Delta^{-1}A$, we obtain posterior samples of the d.r. subspace, denoted as $\{\mathcal{B}^{(1)}, \dots, \mathcal{B}^{(T)}\}$, where T is the number of the posterior samples. If we fix the dimension d then each subspace is a point on the Grassman manifold denoted as $\mathcal{G}_{(d,p)}$, which is the set of all the d dimensional linear subspaces of \mathbb{R}^p . This manifold has a natural Riemannian metric and families of probability distributions can be defined on the Grassmann manifold. See Appendix for details on the differential geometry of this manifold.

The Riemannian metric on the manifold implies the Bayes estimate of the posterior mean should be with respect to the geodesic. This means given subspaces $\{\mathcal{B}^{(1)}, \dots, \mathcal{B}^{(T)}\}$ the posterior summary should be a subspace $\mathcal{B}_{\text{Bayes}}$ that is equidistant to the T posterior samples with respect to the geodesic distance. Given two subspaces \mathcal{W}_1 and \mathcal{W}_2 spanned by orthonormal bases W_1 and W_2 respectively, the geodesic distance between the subspaces is given by the following computation (Karcher, 1977; Kendall, 1990)

$$\begin{aligned} (I - W_1(W_1'W_1)^{-1}W_1')W_2(W_2'W_2)^{-1} &= U\Sigma V' \quad (\text{SVD decomposition}) \\ \Theta &= \text{atan}(\Sigma) \\ \text{dist}(\mathcal{W}_1, \mathcal{W}_2) &= \sqrt{\text{Tr}(\Theta^2)}, \end{aligned}$$

where $\text{Tr}(\cdot)$ is the matrix trace and $\text{atan}(\cdot)$ is the matrix arctangent. Given the above geodesic distance the mean of the subspaces $\{\mathcal{B}^{(1)}, \dots, \mathcal{B}^{(T)}\}$ is the unique subspace with the smallest geodesic distance to the posterior samples

$$\mathcal{B}_{\text{Bayes}} = \arg \min_{\mathcal{B} \in \mathcal{G}(d,p)} \sum_{t=1}^T \text{dist}^2(\mathcal{B}^{(t)}, \mathcal{B}) \quad (2.12)$$

which is called the Karcher mean (Karcher, 1977). We use the algorithm introduced in Absil et al. (2004) to compute the Karcher mean. Given the geodesic distance we can further evaluate the uncertainty of the d.r. subspace by calculating the distances between the mean subspace and the posterior samples. We obtain a standard deviation estimate of the posterior subspace as

$$\text{std}(\{\mathcal{B}^{(1)}, \dots, \mathcal{B}^{(T)}\}) = \sqrt{\frac{1}{T} \sum_{t=1}^T \text{dist}^2(\mathcal{B}^{(t)}, \mathcal{B}_{\text{Bayes}})} \quad (2.13)$$

The posterior distribution on the d.r. subspace is a distribution on the Grassmann man-

ifold $\mathcal{G}_{(d,p)}$. It is of great interest to parameterize and characterize the posterior distribution on this manifold. This is currently beyond the scope of our work.

2.2.4 The $p \gg n$ setting

When the number of predictors is much larger than the sample size, i.e., $p \gg n$, the above procedure is problematic due to the curse of dimensionality. Clustering high dimensional data would be prohibitive due to the lack of samples. This problem can be addressed by slightly adapting computational aspects of the model specification.

Note in our mixture inverse regression model (2.2) and (2.3), μ_{yx} is a mean parameter for $X \mid (Y = y)$, and if $p \gg n$ then it is reasonable to assume that μ_{yx} lies in the subspace spanned by the sample vectors x_1, \dots, x_n – given the limited sample size constraining the d.r. subspace to this subspace is reasonable. By this assumption, $\mu_{yx} - \mu$ and $A\nu_{yx}$, due to equation (2.3), will also be contained in the subspace spanned by the centered sample vectors. Denote \tilde{X} as the $n \times p$ centered predictor matrix, then a singular value decomposition on \tilde{X} yields $\tilde{X} = U_X D_X V_X'$ with the left eigenvectors $U_X \in \mathbb{R}^{n \times p^*}$ and right eigenvectors $V_X \in \mathbb{R}^{p \times p^*}$ where $p^* \leq n \ll p$. In practice one can select p^* by the decay of the singular values. By the above argument for constraints, we can assume $A = V_X \tilde{A}$ with $\tilde{A} \in \mathbb{R}^{p^* \times d}$. We can also assume that $\Delta = V_X \tilde{\Delta} V_X'$ with $\tilde{\Delta} \in \mathbb{R}^{p^* \times p^*}$. The effective number of parameters is thus hugely reduced.

2.2.5 Selecting d

In our analysis the dimension of the d.r. subspace d needs to be determined. In a Bayesian paradigm this is formally a model comparison problem and for two candidate

values d_1 and d_2 the Bayes factor can be used for model selection

$$\text{BF}(d_1, d_2) = \frac{p(\text{data} \mid d_1)}{p(\text{data} \mid d_2)},$$

with the marginal likelihood

$$p(\text{data} \mid d) = \int_{\theta} p(\text{data} \mid d, \theta) p_{\text{prior}}(\theta) d\theta$$

where θ denotes all the relevant model parameters.

The marginal likelihood in our case is obviously not analytically available. Various approximation methods are listed in Lopes and West (2004) yet none of them prove to be computationally efficient in our case due to the existence of the nonparametric prior. We instead adopted out-of-sample validation to select d . For each candidate value d , we obtain a point estimate (the posterior mean) of the d.r. subspace, project out-of-sample test data onto this subspace, and then use the cross-validation error of a predictive model (a classification or regression model) to select d . Empirically this procedure is effective which will be shown in the data analysis.

2.3 Application to simulated and real data

To illustrate the efficacy of BMI we apply it to simulated and real data. The first simulation illustrates how the method captures information on nonlinear manifolds. The second data set is used to compare it to a variety other supervised dimension reduction methods in the classification setting. The third data set illustrates that the method can be used in high-dimensional data.

2.3.1 Regression on a nonlinear manifold

A popular data set used in the manifold learning literature is the swiss roll data, see Figure 2.1. We used the following generative model with $p = 10$ predictors:

$$X_1 = t \cos(t), \quad X_2 = h, \quad X_3 = t \sin(t), \quad X_{4,\dots,10} \stackrel{iid}{\sim} N(0, 1)$$

where $t = \frac{3\pi}{2}(1 + 2\theta)$, $\theta \sim \text{Unif}(0, 1)$, $h \sim \text{Unif}(0, 1)$ and

$$Y = \sin(5\pi\theta) + h^2 + \varepsilon, \quad \varepsilon \sim N(0, 0.01).$$

X_1 and X_3 form an interesting “Swiss roll” shape as illustrated in Figure 2.1(b) and the nonlinear relationship between Y and X_1, X_2, X_3 is illustrated in Figure 2.1 (a). In this case an efficient dimension reduction method should be able to find the first 3 dimensions.

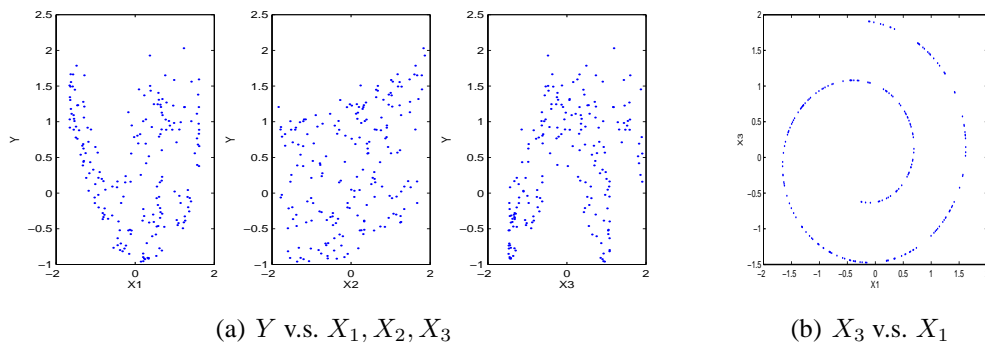


Figure 2.1: Swiss Roll data: Illustration.

For the purpose of comparing methods we used the following metric proposed in Wu et al. (2008) to measure the accuracy in estimating the d.r. space. Let the orthogonal matrix $\hat{B} = (\hat{\beta}_1, \dots, \hat{\beta}_d)$ denote a point estimate of B (which is the first 3 columns of the 10

dimensional identity matrix here), then the accuracy can be measure by

$$\frac{1}{d} \sum_{i=1}^d \|P_B \hat{\beta}_i\|^2 = \frac{1}{d} \sum_{i=1}^d \|(BB') \hat{\beta}_i\|^2$$

where P_B denotes the orthogonal projection onto the column space of B . For BMI \hat{B} is the posterior Karcher mean as proposed in section 2.2.3

We did five experiments corresponding to sample size $n = 100, 200, 300, 400, 500$ from the generative model. In each experiment we applied BMI on 20 randomly drawn datasets with sample size n and averaged the accuracies measured as stated above. For BMI we ran 10000 MCMC iterations and used a burn-in of 5000 and set $d = 3$ and used the Gaussian kernel in (2.10). Figure 2.2 shows the performance of BMI as well as that by a variety of SDR methods: SIR (Li, 1991), local sliced inverse regression LSIR (Wu et al., 2008), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) and principal Hessian directions (pHd) (Li, 1992). The accuracies for SIR, LSIR, SAVE and PHd are copied from Wu et al. (2008) except for the scenario of $n = 100$. It is clear that BMI consistently has the best accuracy. LSIR is the most competitive of the other methods as one would expect since it shares with BMI the idea of localizing the inverse regression around a mixture or partition.

Of particular interest is the estimate uncertainty. As stated in section 2.2.3 the Karcher mean (2.12) of the posterior samples is taken as a point estimate, and a natural uncertainty measure is simply the standard deviation as defined in (2.13). For illustration we applied our method on a data set with sample size 400. Figure 2.3 shows a boxplot for the distances between the posterior sampled subspaces and the posterior Karcher mean subspace and the standard deviation is calculated to be 0.2162. It is also calculated that the distance between the Karcher mean and the true d.r. subspace is 0.2799. It is interesting the see that the true d.r. subspace lies “not far” (compared with the standard deviation) from our point estimate.

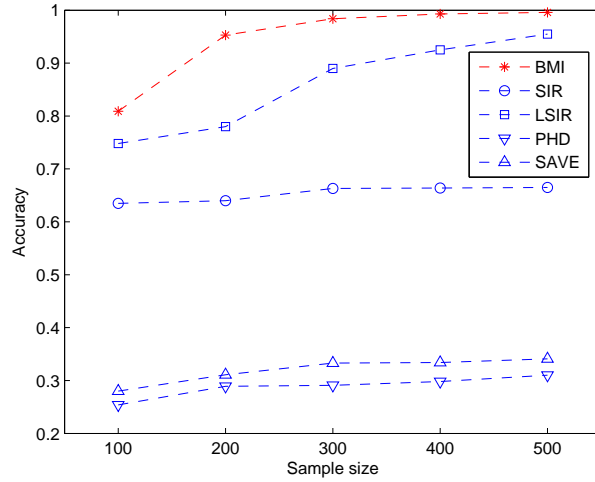


Figure 2.2: Swiss Roll data: Accuracy for different methods.

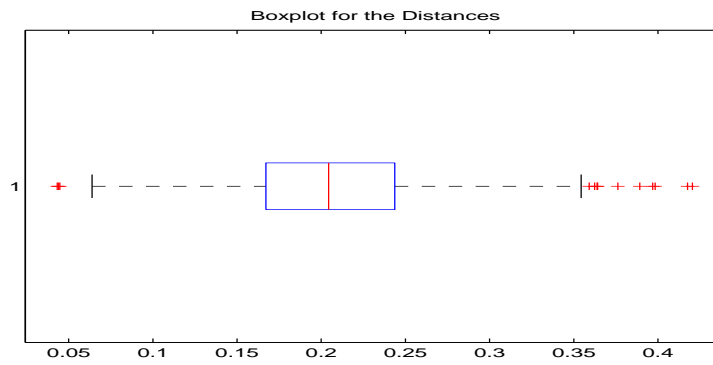


Figure 2.3: The boxplot for the distances between the posterior samples and their Karcher mean.

To illustrate that BMI is borrowing of information across the response variables we plot in Figure 2.4 the cluster labels for all the samples as ordered in terms of the magnitude of response for the last MCMC iteration in an experiment with sample size $n = 200$. Samples with similar responses tend to be clustered and have similar underlying clustering distributions which change "gradually" with increasing response instead of "rigidly" as what would be obtained by the slicing procedure in SIR.

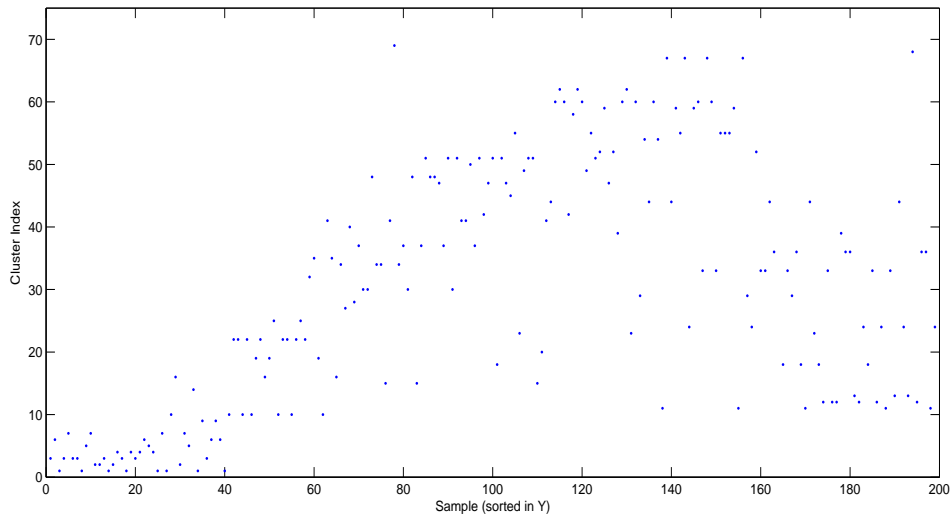


Figure 2.4: Cluster labels for all the 200 samples at the last iteration under an experiment. The samples are ordered in terms of the magnitude of Y . Closer samples (in terms of Y) seem to have similar underlying clustering distributions which change "gradually" with increasing response.

We utilized cross-validation to select the number of d.r. directions d in a case of sample size 200. For each value of $d \in \{1, \dots, 10\}$, we project out-of-sample data onto the d -dimensional space and a nonparametric kernel regression model to predict the response. The error reported is the mean square prediction error. The error v.s. different candidate values of d is depicted in Figure 2.5. The smallest error corresponds to $d = 3$, the true number of d.r. directions.

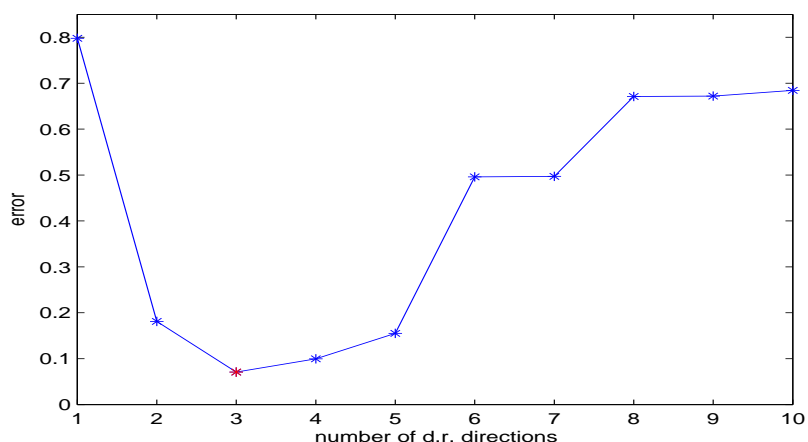


Figure 2.5: Swiss Roll data: Error v.s. number of d.r directions kept. The minimum one corresponds to $d = 3$, the true value.

2.3.2 Classification

In Sugiyama (2007) a variety of SDR methods were compared on the Iris data set available from the UCI machine learning repository ¹, originally from Fisher (1936). The data consists of 3 classes with 50 instances of each class. Each class refers to a type of Iris plant (“Setosa”, “Virginica” and “Versicolour”), and has 4 predictors describing the length and width of the sepal and petal. The methods compared in Sugiyama (2007) were Fisher’s linear discriminant analysis (FDA), local Fisher discriminant analysis (LFDA) (Sugiyama, 2007), locality preserving projections (LPP) (He and Niyogi, 2003), LDI (Hastie and Tibshirani, 1996a), neighbourhood component analysis (NCA) (Goldberger et al., 2005), and metric learning by collapsing classes (MCML) (Globerson and Roweis, 2006).

To demonstrate that BMI can find multiple clusters we merge “Setosa”, “Virginica” into a single class and examine whether we are able to separate them.

In Figures 2.6 we plot the projection of the data onto a 2 dimensional d.r. subspace. We set $\alpha_0 = 1$ in (2.7). The classes are separated as are the two clusters in the merged “Setosa”,

¹<http://archive.ics.uci.edu/ml/datasets/Iris>

“Virginica” class. Our method is able to further embed the data into a 1 dimensional d.r. subspace while still preserving the separation structure (Figure 2.7). Figure 2.8 is a copy of the Figure 6 in Sugiyama (2007) and provides a comparison of FDA, LFDA, LPP, LDI, NCA, and MCML. Comparing Figure 2.6 and 2.7 with Figure 2.8 we see that BMI and NCA are similar with respect to performance, and they both have the advantage of being able to embed this particular data into a 1 dimensional d.r. subspace while the others cannot.

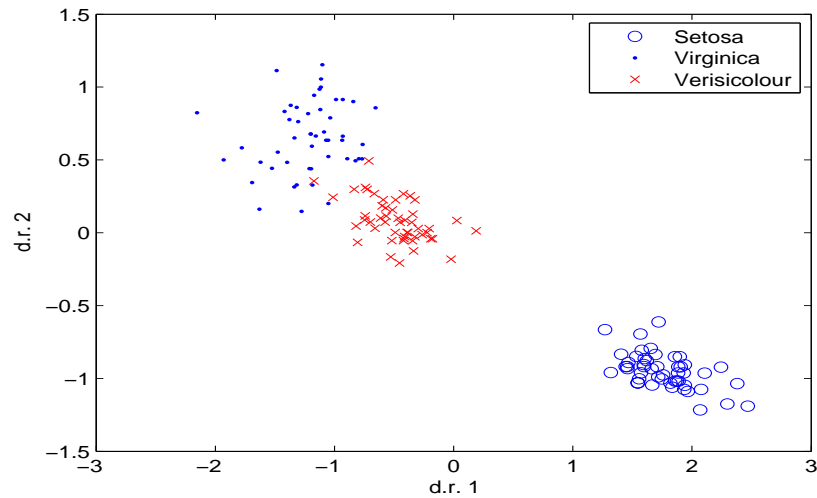


Figure 2.6: Visualization of the embedded *Iris* data onto a 2 dimension subspace.

2.3.3 High-dimensional data: digits

The MNIST digits data² is commonly used in the machine learning literature to compare algorithms for classification and dimension reduction. The data set consists of 60,000 images of handwritten digits, $\{0, 1, \dots, 9\}$ where each image is considered as a vector of $28 \times 28 = 784$ gray-scale pixel intensities. The utility of the digits data is that the d.r. directions have a visually intuitive interpretation.

We apply BMI to two binary classification tasks: digits 3 v.s. 8, and digits 5 v.s. 8.

²<http://yann.lecun.com/exdb/mnist/>

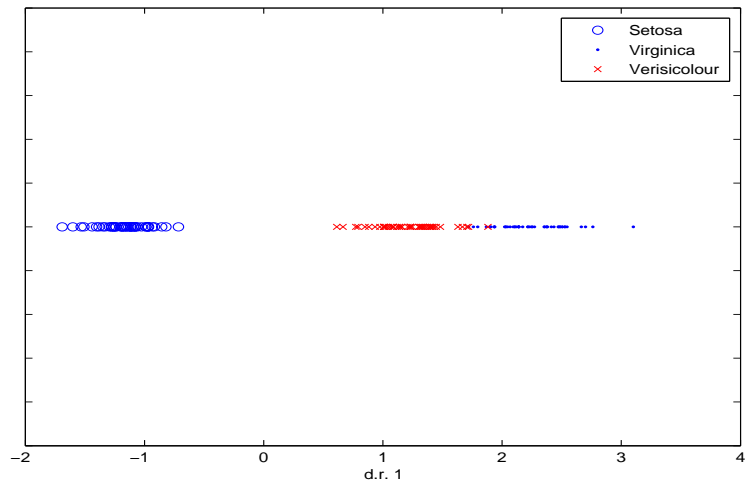


Figure 2.7: Visualization of the embedded *Iris* data onto a 1 dimensional subspace.

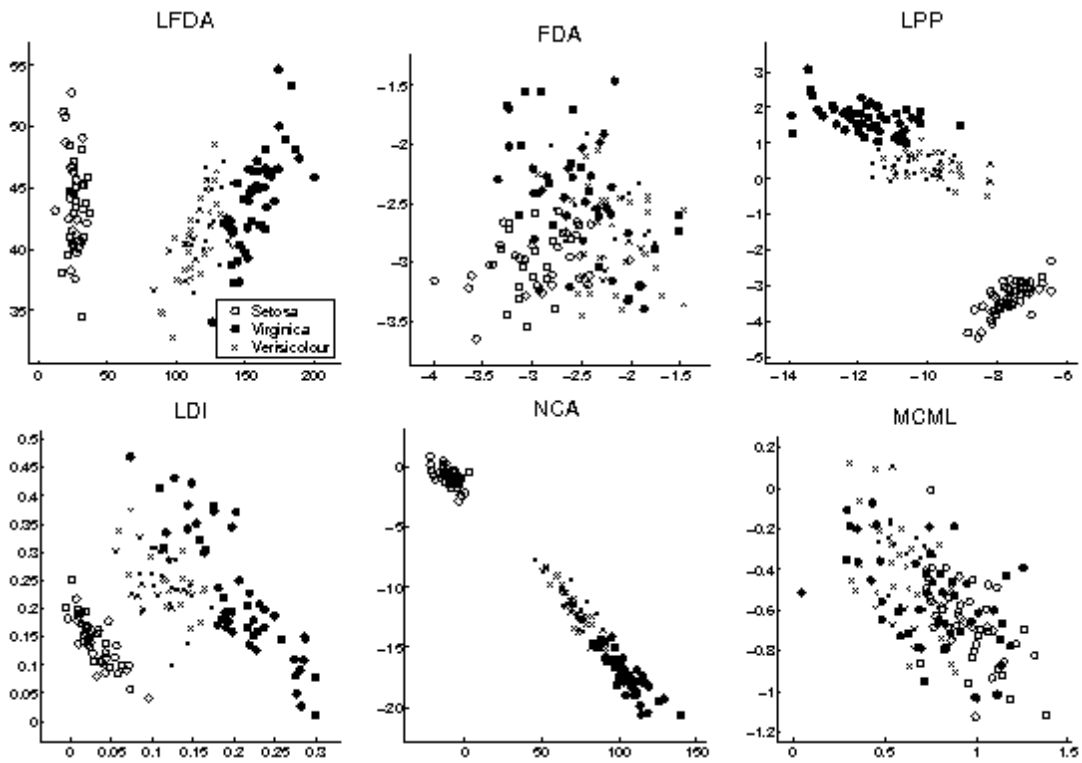


Figure 2.8: Visualization of the *Iris* data for different methods. (see also Sugiyama (2007) Figure 6.)

In each task we randomly select 200 images, 100 for each digit. Since the number of predictors is far greater than the sample size ($p \gg n$), we used the modification of BMI described in Section 2.2.4 and $p^* = 30$ eigenvectors are selected. We run BMI for 10000 iterations with the first as 5000 burn-in and choose $d = 1$. The posterior means of the top d.r. direction, depicted in a 28×28 pixel format, are displayed in Figures 2.9. We see that the top d.r. directions precisely capture the difference between digits 3 and 8, an upper left and lower left region, and the difference between digits 5 and 8, an upper right and lower left region.

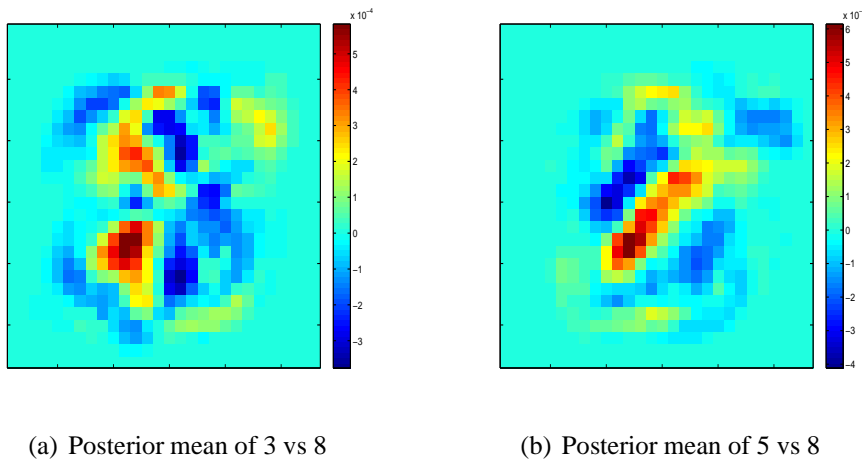


Figure 2.9: BMI: (a) The posterior mean of the top d.r. direction for 3 versus 8, shown in a 28×28 pixel format. (b) The posterior mean of the top d.r. direction for 5 versus 8, shown in a 28×28 pixel format. Difference between digits is reflected by the red color.

2.4 Discussion and Future Work

We have proposed a Bayesian framework for supervised dimension reduction using a highly flexible nonparametric Bayesian mixture modeling approach. Our method retrieves the dimension reduction or d.r. subspace by the utilization of the dependent Dirichlet process that allows for natural clustering for the data in terms of both the response and predictor variables.

Our BMI model highlights a flexible generalization of the PFC framework to a non-parametric setting and addresses the issue of multiple clusters for a slice of the response. This idea of multiple clusters suggests that this approach is relevant even when the marginal distribution of the predictors is not concentrated on a linear subspace. The idea of modeling nonlinear subspaces is central in the area of manifold learning (Roweis and Saul, 2000; Tenenbaum et al., 2000; Donoho and Grimes, 2003). BMI is one probabilistic formulation of a supervised manifold learning algorithm.

A fundamental issue raised by this methodology is the development of distribution theory on the Grassmann manifold. There has been work on uniform distributions on the Grassmann manifold and we discuss the case corresponding to subspaces drawn from a fixed number of centered normals. To better characterize the uncertainty of our posterior estimates it would be of great interest to develop richer distributions on the Grassmann manifold. In addition it will be of great interest to directly develop proposal distributions for the d.r. subspace on the Grassmann manifold so that inference could be more coherent.

A recently proposed unsupervised framework called "(probabilistic) local dimension reduction" allows different linear dimension reduction for data in different region. For example, in Chen et al. (2009) a factor model with mixtures on the "loading matrix" is proposed

$$X \sim N(\mu + A_X w, \phi^{-1}I)$$

where the subscript X is added to A to demonstrate mixtures on the loading matrix. This is meaningful because by allowing different linear dimension reduction in different data region the local Euclidean structure can be captured, hence the underlying manifold structure could be successfully learned. It will be of great interest to extend this framework to supervised settings to incorporate the information in the response in determining the local dimension reduction.

Chapter 3

Bayesian Gradient Learning Through Kernel Models for Supervised Dimension Reduction

3.1 Introduction

As detailed in Section 1.3.3 supervised dimension reduction methods can be divided into three categories: methods based on forward regression, methods based on learning gradients of the regression function, and methods based on inverse regression. In Chapter 2 we develop a highly flexible nonparametric Bayesian mixture model that falls into the category of inverse regression. In this chapter we focus on the gradient learning category.

The underlying model in supervised dimension reduction is given in (1.6). A slightly restricted framework in which the reduction is captured in the regression (mean) function is also commonly adopted: denote the regression function $f(x) = E(Y|X = x)$, then the additive error model

$$Y = f(X) + \varepsilon = \tilde{g}(b'_1 X, \dots, b'_d X) + \varepsilon, \quad (3.1)$$

defines $B = (b_1, \dots, b_d)$ as the d.r. space that reduces the original X to $b'_1 X, \dots, b'_d X$ in capturing the regression function.

There are a number of SDR methods that belong to the gradient learning category. Minimum average variance estimation (MAVE) (Xia et al., 2002) retrieves predictive directions by eigen-decomposition of the Gradient Outer Product matrix (GOP) $E(\nabla f \nabla f')$ where the gradient ∇f is estimated by local polynomial fitting. However the polynomial

fitting adopted by MAVE cannot be used directly when the number of predictor variables p is larger than the sample size n due to over-fitting and numerical instability. The methods proposed in Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Mukherjee et al. (2006); Wu et al. (2007) also focus on the GOP matrix $E(\nabla f \nabla f')$ yet uses kernel methods to learn the gradient and overcome the over-fitting problems in the $p > n$ scenario by adding a regularization term in the gradient estimate. The utilization of kernels introduces nonlinear features to the problem hence makes these methods more flexible. The main drawback, however, is their lack of probabilistic interpretations.

We formulate a coherent Bayesian nonparametric model for supervised dimension reduction that is based on ideas from the above methods and adds appropriate probabilistic and statistical frameworks that allow for uncertainty estimation which is important for measuring the estimates. We will illustrate how our Bayesian model allows for formal inference of uncertainty in dimension reduction as well as inference the uncertainty of conditional dependencies in graphical models.

In Section 3.2 we state a statistical basis for dimension reduction and state the learning gradients approach developed in Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Mukherjee et al. (2006); Wu et al. (2007). In Section 3.3 we develop a fully Bayesian nonparametric model for learning gradients and provide a Markov chain Monte Carlo procedure for inference of model parameters. In Section 3.4 we illustrate the method and address questions about mixing of the MCMC using simulated data and present analysis on real data. We close with a short discussion.

3.2 Dimension reduction and conditional independence based on gradients

3.2.1 Euclidean setting

The central quantity of interest here is the gradient outer product (GOP) matrix. We first formulate its properties in the Euclidean p -dimensional ambient space. Assume the regression function $f(x)$ is smooth, the gradient is given by $\nabla f = (\frac{\partial f}{\partial x^1}, \dots, \frac{\partial f}{\partial x^p})'$ and the the gradient outer product matrix Γ is a $p \times p$ defined as

$$\Gamma = E_X [(\nabla f) (\nabla f)']. \quad (3.2)$$

The relation between the gradient outer product matrix and dimension reduction is illustrated by the following observation (Wu et al., 2007). Under the assumptions of the semi-parametric model (3.1), the gradient outer product matrix Γ is of rank at most d and if we denote by $\{v_1, \dots, v_d\}$ the eigenvectors associated to the nonzero eigenvalues of Γ then following holds

$$\text{span}(B) = \text{span}(v_1, \dots, v_d)$$

The GOP matrix Γ is related with the covariance of the inverse regression matrix $\Omega_{X|Y} = \text{cov}[E(X|Y)]$ developed in Li (1991). In Wu et al. (2007) it was shown that for a linear regression function

$$\Gamma = \sigma_Y^2 \left(1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}\right)^2 \Sigma_X^{-1} \Omega_{X|Y} \Sigma_X^{-1} \quad (3.3)$$

where σ_Y^2 is the variance of the response variable, σ_ε^2 is the variance of the error and $\Sigma_X = \text{cov}(X)$ is the covariance matrix of the predictor variables. This suggests that the GOP matrix Γ is the transformed covariance matrix of the inverse regression function $\Omega_{X|Y}$. For nonlinear regression functions that are smooth, one can partition the predictor space

such that in each partition the regression function is approximately linear hence (3.3) holds locally. Then the GOP can be regarded as the weighted sum of the local transformed covariance matrix of the inverse regression function (Wu et al., 2007).

3.2.2 Manifold setting

The above statements are formulated with respect to Euclidean geometry or linear subspaces. The local nature of the gradient allows for an interpretation of the gradient outer product in the manifold setting (Wu et al., 2007; Mukherjee et al., 2006). In the manifold setting, the support of marginal measure of the predictor variables is concentrated on a manifold \mathcal{M} of dimension $d_{\mathcal{M}} \ll p$. We assume the existence of an isometric embedding from the manifold to the ambient space, $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$. The observed predictor variables are the image of points drawn from a distribution concentrated on the manifold, $x_i = \varphi(q_i)$ where $(q_i)_{i=1}^n$ are points concentrated on the manifold. The gradient on the manifold $\nabla_{\mathcal{M}}f$ is a well defined mathematical quantity and is a $d_{\mathcal{M}}$ -dimensional vector. However, since we only obtain observations in the p -dimensional ambient space we cannot compute the gradient on the manifold. Given data one can estimate the gradient in the ambient space, \vec{f} , which is a p -dimensional vector. It was shown in Mukherjee et al. (2006) that under weak conditions on the manifold and regression an estimate of the gradient in the ambient space, \vec{f} , is consistent in the following sense

$$(d\varphi)^* \vec{f} \longrightarrow \nabla_{\mathcal{M}}f \quad \text{as} \quad n \rightarrow \infty,$$

where $(d\varphi)^*$ is the dual of $d\varphi$. This suggests that the gradient estimate in the ambient space provides information on the gradient on the manifold.

3.2.3 Conditional independence

The theory of Gauss-Markov graphs (Speed and Kiiveri, 1986; Lauritzen, 1996) was developed for multivariate Gaussian densities

$$p(x) \propto \exp\left(-\frac{1}{2}x^T J X + h^T x\right),$$

where the covariance is J^{-1} and the mean is $\mu = J^{-1}h$. The result of the theory is that the precision matrix J provides a measurement of conditional independence. The meaning of this dependence is highlighted by the partial correlation matrix R where each element r_{ij} is a measure of dependence between variables i and j conditioned on all other variables $S^{/ij}$ and $i \neq j$

$$r_{ij} = \frac{\text{cov}(x_i, x_j | S^{/ij})}{\sqrt{\text{var}(x_i | S^{/ij})} \sqrt{\text{var}(x_j | S^{/ij})}} = -\frac{J_{ij}}{\sqrt{J_{ii}J_{jj}}}.$$

Under the assumptions implied by equations (3.2) and (3.3) the gradient outer product matrix Γ is a covariance matrix. So we can apply the theory of Gauss-Markov graphs to Γ and consider the matrix $J_\Gamma = \Gamma^{-1}$. The advantage of computing this matrix in the regression and classification framework is that it provides an estimate of the conditional dependence of the predictor variables with respect to variation of the response variable. The modeling assumption of our construction is that the matrix J_Γ is sparse with respect to the factors or directions (b_1, \dots, b_d) rather than the p predictor variables. Under this assumption we use pseudo-inverses in order to construct the dependence graph based on the partial correlation R_Γ .

3.3 Bayesian Gradient Learning Through Nonparametric Kernel Models

Many approaches for the inference of gradients exist including various numerical derivative algorithms, local linear smoothing, and learning gradients by kernel models (Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006). Our approach will be closely related to the penalized likelihood or regularization models developed in Mukherjee and Zhou (2006); Mukherjee and Wu (2006).

3.3.1 The model

The starting point for our gradient estimation procedure is the first order Taylor series expansion of the regression function $f(x)$ around a point u

$$f(x) = f(u) + \nabla f(x)'(x - u) + \varepsilon_d, \quad (3.4)$$

where the deterministic error term $\varepsilon_d = O(\|x - u\|^2)$ is a function of the distance between x and u and the model

$$y = f(x) + \varepsilon, \quad (3.5)$$

where ε models the stochastic noise. For simplicity we work with a fixed design model with $(x_i)_{i=1}^n$ given (see (Liang et al., 2007) for the development of the random design setting of which this is a special case). Coupling equations (3.4) and (3.5) we can state the following model

$$y_i = \frac{1}{n} \left[\sum_{j=1}^n f(x_j) + \vec{f}(x_i)'(x_i - x_j) + \varepsilon_{ij} \right], \quad \text{for } i = 1, \dots, n, \quad (3.6)$$

$$\varepsilon_{ij} = y_i - f(x_j) - \vec{f}(x_i)'(x_i - x_j), \quad \text{for } i, j = 1, \dots, n \quad (3.7)$$

where f models the regression function, \vec{f} models the gradient, and ε_{ij} has both stochastic and deterministic components varying monotonically as a function of the distance between two observations x_i and x_j . We will model ε_{ij} as a random quantity and use a simple spatial model to specify the covariance structure. Specifically, we first define an association matrix with $w_{ij} = \exp(-\|x_i - x_j\|^2/2s^2)$ with fixed bandwidth parameter s . We then define $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, (\phi/w_{ij})^{-1})$ where ϕ will be a random scale parameter. Define the vector $\varepsilon_{i\bullet} = (\varepsilon_{i1}, \dots, \varepsilon_{in})'$, a joint probability density function on this vector can be used to specify a likelihood function for the data. We specify the following model for $\varepsilon_{i\bullet}$.

$$p(\varepsilon_{i\bullet}) \propto \phi^{\frac{n}{2}} \exp \left\{ -\frac{\phi}{2} (\varepsilon_{i\bullet}' W_i \varepsilon_{i\bullet}) \right\}, \quad (3.8)$$

where the diagonal matrix $W_i = \text{diag}(w_{i1}, \dots, w_{in})$.

The key here is to appropriately model f and \vec{f} . In Chapter 4 we build nonparametric kernel models in a Bayesian framework for learning unknown functions. Through a proper modeling of the Reproducing Kernel Hilbert Space (RKHS) (Wahba, 1990) generated by a kernel function $K(\cdot, \cdot)$ using Dirichlet process priors (Ferguson, 1973, 1974; Müller et al., 2004; MacEachern and Müller, 1998; Escobar and West, 1995; Sethuraman, 1994), we obtain kernel representation forms for the unknown functions (see (4.4) in Section 4.2.3), analogous to the “representer theorem” in Kimeldorf and Wahba (1971). See Chapter 4 for details. Specifically we could apply nonparametric kernel models for the regression function and gradient function and obtain:

$$f(x) = \sum_{i=1}^n \alpha_i K(x, x_i), \quad \vec{f}(x) = \sum_{i=1}^n \mathbf{c}_i K(x, x_i) \quad (3.9)$$

where $\alpha = (\alpha_1, \dots, \alpha_n)' \in \mathbb{R}^n$, $C = (\mathbf{c}_1, \dots, \mathbf{c}_n) \in \mathbb{R}^{p \times n}$.

Substituting the above representation in equation (3.7) results in the following parame-

trized model

$$y_i = \sum_{k=1}^n \alpha_k K(x_j, x_k) + \sum_{k=1}^n (\mathbf{c}'_k(x_i - x_j)) K(x_i, x_k) + \varepsilon_{ij}, \quad \text{for } i, j = 1, \dots, n. \quad (3.10)$$

We can rewrite the above in matrix notation where for the i -th observation

$$y_i \mathbf{1} = K \alpha + D_i C K_i + \varepsilon_{i\bullet},$$

where $\mathbf{1}$ is the $n \times 1$ vector of all 1's, K_i is the i -th column of the gram matrix K where $K_{ij} = k(x_i, x_j)$, E is the $n \times p$ data matrix and $D_i = \mathbf{1}x'_i - E$.

Direct prior specification tends to be problematic since this model has a huge number of parameters, for instance C itself has $p \times n$ parameters. West (2003) defined and exemplified the use of a flexible class of generalized g-priors that allows for different degrees of shrinkage estimation of regression parameters in different principal component directions on the induced design space for a regression model and proves to be very effective in problems with many parameters and relatively small sample size, a feature typically termed “large p and small n”. In West (2003) a generalized g-prior is induced by independent normal priors on the regression parameters of the equivalent principal component regression transformation of the model. Applying this strategy in our setting leads to spectral decompositions to reduce the number of variables:

- Let $M_X = (x_1 - x_n, \dots, x_{n-1} - x_n) \in \mathbb{R}^{p \times (n-1)}$ be the sample difference matrix, so that $\text{rank}(M_X) = d_x \leq \min(n-1, p)$. Let the singular value decomposition of M_X be $M_X = V_M \Lambda_M U'_M$ s.t. $V'_M V_M = I_{d_x}$ and $\Lambda_M \in \mathbb{R}^{d_x \times d_x}$ is diagonal with the singular values of M_X in a descending order. Ignoring small singular values yields $M_X = \tilde{V} \tilde{\Lambda} \tilde{U}'$ where $\tilde{V} \in \mathbb{R}^{p \times d^*}$ is the first d^* columns of V_M . Since each row vector of D_i lies in the column space of M_X hence the column space of \tilde{V} , then

$\exists \tilde{D}_i \in \mathbb{R}^{n \times d^*}$ s.t. $D_i = \tilde{D}_i \tilde{V}'$. Put $\tilde{C} = \tilde{V}' C$, then $D_i C = \tilde{D}_i \tilde{V}' C = \tilde{D}_i \tilde{C}$.

- A spectral decomposition can also be applied to the gram matrix K resulting in $K = F_K \Lambda_K F_K'$. Note that $K\alpha = F\beta$ where $F = F_K \Lambda_K$, $\beta = F_K' \alpha$. We can again select columns of F corresponding to the largest eigenvalues m .

Given the above re-parametrization we have for the i -th observation

$$y_i \mathbf{1} = F\beta + \tilde{D}_i \tilde{C} K_i + \varepsilon_{i\bullet}.$$

The prior specifications are now on the parameters (ϕ, β, \tilde{C}) :

$$\phi \propto \frac{1}{\omega},$$

$$\beta \sim N(0, \Delta_\psi^{-1}) \text{ where } \Delta_\psi = \text{diag}(\psi_1, \dots, \psi_m) \text{ and } \psi_i \sim \text{Gamma}(a_\psi/2, b_\psi/2),$$

$$\tilde{C}_j \sim N(0, \Delta_\varphi^{-1}) \text{ where } \Delta_\varphi = \text{diag}(\varphi_1, \dots, \varphi_{d^*}) \text{ and } \varphi_i \sim \text{Gamma}(a_\varphi/2, b_\varphi/2),$$

where \tilde{C}_j is the j -th column of \tilde{C} and $a_\psi, b_\psi, a_\varphi, b_\varphi, \omega$ are pre-specified hyper-parameters, and an improper prior for ϕ is used. These priors imply generalized g-priors (West, 2003) on the original parameters α and C .

Given the probability model for the error vector in (3.8), the likelihood of our model given observations $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ is

$$\text{Lik}(D|\phi, \beta, \tilde{C}) \propto \phi^{\frac{n^2}{2}} \exp \left\{ -\frac{\phi}{2} \sum_{i=1}^n \left(y_i \mathbf{1} - F\beta - \tilde{D}_i \tilde{C} K_i \right)' W_i \left(y_i \mathbf{1} - F\beta - \tilde{D}_i \tilde{C} K_i \right) \right\}, \quad (3.11)$$

where the diagonal matrix $W_i = \text{diag}(w_{i1}, \dots, w_{in})$.

3.3.2 Sampling from the posterior

A standard Gibbs sampler can be used to simulate the posterior density, $\text{Post}(\phi, \beta, \tilde{C}|D)$, due to the normal form of the likelihood and conjugacy properties of the prior specifications. The update steps of the Gibbs sampler given data D and initial values $(\phi^{(0)}, \beta^{(0)}, \tilde{C}^{(0)})$ follow:

1. Update Δ_ψ : $\Delta_\psi^{(t+1)} = \text{diag}(\psi_1^{(t+1)}, \dots, \psi_m^{(t+1)})$ with

$$\psi_i^{(t+1)} | D, \phi^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} \sim \text{Gamma} \left(\frac{a_\psi + 1}{2}, \frac{b_\psi + (\beta_i^{(t)})^2}{2} \right), \quad i = 1, \dots, m$$

where $\beta_i^{(t)}$ is the i -th element of $\beta^{(t)}$;

2. Update Δ_φ :

$$\Delta_\varphi^{(t+1)} = \text{diag}(\varphi_1^{(t+1)}, \dots, \varphi_{d^*}^{(t+1)})$$

$$\varphi_i^{(t+1)} | D, \phi^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} \sim \text{Gamma} \left(\frac{a_\varphi + 1}{2}, \frac{b_\varphi + \sum_{j=1}^n (\tilde{C}_{ij}^{(t)})^2}{2} \right), \quad i = 1, \dots, d^*,$$

where $\tilde{C}_{ij}^{(t)}$ is the (i, j) -th element of $\tilde{C}^{(t)}$;

3. Update β :

$$\beta^{(t+1)} | D, \tilde{C}^{(t)}, \Delta_\psi^{(t+1)}, \phi^{(t)} \sim N(\mu_\beta, \Sigma_\beta)$$

with

$$\Sigma_\beta = \left(F' \left(\sum_{i=1}^n \phi^{(t)} W_i \right) F + \Delta_\psi^{(t+1)} \right)^{-1},$$

$$\mu_\beta = \phi^{(t)} \Sigma_\beta F' \sum_{i=1}^n W_i (y_i \mathbf{1} - \tilde{D}_i \tilde{C}^{(t)} K_i);$$

4. Update $\tilde{C} = (\tilde{C}_1, \dots, \tilde{C}_n)$:

For \tilde{C}_j with $j = 1, \dots, n$

$$\tilde{C}_j^{(t+1)} | D, \tilde{C}_{\setminus j}^{(t)}, \Delta_\psi^{(t+1)}, \phi^{(t)} \sim N(\mu_j, \Sigma_j),$$

where $\tilde{C}_{\setminus j}^{(t)}$ is the matrix $\tilde{C}^{(t)}$ with the j -th column removed.

$$\begin{aligned} b_{ij} &= y_i \mathbf{1} - F \beta^{(t+1)} - \tilde{D}_i \sum_{k \neq j} \tilde{C}_k^{(t)} K_{ik} \\ \Sigma_{j0} &= \left(\phi^{(t)} \sum_{i=1}^n K_{ij}^2 \tilde{D}_i' W_i \tilde{D}_i \right)^{-1}, \\ \mu_{j0} &= \phi^{(t)} \Sigma_j \sum_{i=1}^n K_{ij} \tilde{D}_i' W_i b_{ij} \\ \Sigma_j &= (\Sigma_{j0}^{-1} + \Delta_\varphi^{(t+1)})^{-1}, \\ \mu_j &= \Sigma_j (\Sigma_{j0}^{-1} \mu_{j0}). \end{aligned}$$

5. Update ϕ :

$$\phi^{(t+1)} | D, \tilde{C}^{(t+1)}, \beta^{(t+1)} \sim \text{Gamma}(a, b),$$

where

$$\begin{aligned} a &= \frac{n^2}{2}, \\ b &= \frac{1}{2} \left(\sum_{i=1}^n \left[y_i \mathbf{1} - F \beta^{(t+1)} - \tilde{D}_i \tilde{C}^{(t+1)} K_i \right]' W_i \left[y_i \mathbf{1} - F \beta^{(t+1)} - \tilde{D}_i \tilde{C}^{(t+1)} K_i \right] \right). \end{aligned}$$

3.3.3 Posterior Inference

Given draws $\{\tilde{C}^{(t)}\}_{t=1}^T$ from the posterior distribution we can compute $\{C^{(t)}\}_{t=1}^T$ from the relation $\tilde{C} = \tilde{V}'C$ which allows us to compute posterior draws of the GOP matrix

$$\Gamma_D^{(t)} = C^{(t)} K K' (C^{(t)})'$$

We can then compute the posterior mean GOP matrix as well as a variance estimate

$$\hat{\mu}_{\Gamma, D} = \frac{1}{T} \sum_{t=1}^T \Gamma_D^{(t)}, \quad \hat{\sigma}_{\Gamma, D} = \frac{1}{T} \sum_{t=1}^T (\Gamma_D^{(t)} - \hat{\mu}_{\Gamma, D})_e^2$$

where $(\cdot)_e^2$ denotes the element-wise square.

As formulated in Section 2.2.3 the d.r. subspace is on the manifold $\mathcal{G}_{(d,p)}$ which is the set of all the d dimensional linear subspaces of \mathbb{R}^p and the Riemannian metric on the manifold implies the Bayes estimate of the posterior mean should be with respect to the geodesic. A posterior distribution on this manifold is naturally induced by the posterior distribution of the gradient. A spectral decomposition of $\Gamma_D^{(t)}$ keeping the eigen-vectors corresponding to the largest d eigen-values provides a posterior draw of the d.r. subspace $\mathcal{B}^{(t)}$, and the posterior Karcher mean (Karcher, 1977) as well as standard deviation of the d.r. subspace can be readily computed by (2.12) and (2.13).

For inference of conditional independence we compute the conditional independence and partial correlation matrices

$$J^{(t)} = (\Gamma_D^{(t)})^{-1}, \quad R_{ij}^{(t)} = -\frac{J_{ij}^{(t)}}{\sqrt{J_{ii}^{(t)} J_{jj}^{(t)}}},$$

using a pseudo-inverse to compute $(\Gamma_D^t)^{-1}$. The mean and variance of the posterior esti-

mates of the partial correlations can be computed as above using $\{R^{(t)}\}_{t=1}^T$

$$\hat{\mu}_{R,D} = \frac{1}{T} \sum_{t=1}^T R^{(t)}, \quad \hat{\sigma}_{R,D} = \frac{1}{T} \sum_{t=1}^T (R^{(t)} - \hat{\mu}_{R,D})^2$$

These quantities could be used to infer a graphical model with the capability to evaluate the uncertainty of the correlation structure.

3.3.4 Binary regression

The extension to classification problems where responses are $y_i = 1/0$ using a probit link function is implemented using a set of latent variables $Z = (z_1, \dots, z_n)'$ modeled as a truncated normal distribution with standard variance. In this setting $\phi \equiv 1$ and the same Gibbs sampler with a step added to sample the latent variable can be used to sample from the posterior density, $\text{Post}(\beta, \tilde{C}|D)$. The update steps of the Gibbs sampler given data D and initial values $(Z^{(0)}, \beta^{(0)}, \tilde{C}^{(0)})$ follow:

1. Update Δ_ψ : $\Delta_\psi^{(t+1)} = \text{diag}(\psi_1^{(t+1)}, \dots, \psi_m^{(t+1)})$ with

$$\psi_i^{(t+1)} | D, Z^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} \sim \text{Gamma} \left(\frac{a_\psi + 1}{2}, \frac{b_\psi + (\beta_i^{(t)})^2}{2} \right), \quad i = 1, \dots, m$$

2. Update Δ_φ :

$$\Delta_\varphi^{(t+1)} = \text{diag}(\varphi_1^{(t+1)}, \dots, \varphi_{d^*}^{(t+1)})$$

$$\varphi_i^{(t+1)} | D, Z^{(t)}, \beta^{(t)}, \tilde{C}^{(t)} \sim \text{Gamma} \left(\frac{a_\varphi + 1}{2}, \frac{b_\varphi + \sum_{j=1}^n (\tilde{C}_{ij}^{(t)})^2}{2} \right), \quad i = 1, \dots, d^*,$$

where $\tilde{C}_{ij}^{(t)}$ is the (i, j) -th element of $\tilde{C}^{(t)}$;

3. Update β :

$$\beta^{(t+1)} | D, \tilde{C}^{(t)}, \Delta_\psi^{(t+1)}, Z^{(t)} \sim N(\mu_\beta, \Sigma_\beta)$$

with

$$\begin{aligned}\Sigma_\beta &= \left(F' \left(\sum_{i=1}^n W_i \right) F + \Delta_\psi^{(t+1)} \right)^{-1}, \\ \mu_\beta &= \Sigma_\beta F' \sum_{i=1}^n W_i (z_i^{(t)} \mathbf{1} - \tilde{D}_i \tilde{C}^{(t)} K_i); \end{aligned}$$

4. Update $\tilde{C} = (\tilde{C}_1, \dots, \tilde{C}_n)$:

For \tilde{C}_j with $j = 1, \dots, n$

$$\tilde{C}_j^{(t+1)} | D, \tilde{C}_{\setminus j}^{(t)}, \Delta_\psi^{(t+1)}, Z^{(t)} \sim N(\mu_j, \Sigma_j),$$

where $\tilde{C}_{\setminus j}^{(t)}$ is the matrix $\tilde{C}^{(t)}$ with the j -th column removed.

$$\begin{aligned} b_{ij} &= z_i^{(t)} \mathbf{1} - F \beta^{(t)} - \tilde{D}_i \sum_{k \neq j} \tilde{C}_k^{(t)} K_{ik} \\ \Sigma_{j0} &= \left(\sum_{i=1}^n K_{ij}^2 \tilde{D}_i' W_i \tilde{D}_i \right)^{-1}, \\ \mu_{j0} &= \Sigma_j \sum_{i=1}^n K_{ij} \tilde{D}_i' W_i b_{ij} \\ \Sigma_j &= (\Sigma_{j0}^{-1} + \Delta_\varphi^{(t+1)})^{-1}, \\ \mu_j &= \Sigma_j (\Sigma_{j0}^{-1} \mu_{j0}). \end{aligned}$$

5. Update Z :

For $i = 1, \dots, n$

$$z_i^{(t+1)} | D, \beta^{(t+1)}, \tilde{C}^{(t+1)} \sim \begin{cases} N^+(\eta_i, 1) & \text{for } y_i = 1 \\ N^-(\eta_i, 1) & \text{for } y_i = 0 \end{cases}$$

where N^+ and N^- denote the positive and negative truncated normal distributions

and $(\eta_1, \dots, \eta_n)' = F\beta^{(t+1)}$.

3.3.5 Selecting d

The decision of how many d.r. directions to keep can in theory rely upon the posterior distribution of the eigenvalues of the gradient outer product matrices drawn by simulating from the posterior. In practice we can use the cross-validation procedure formulated in Section 2.2.5.

3.3.6 Modeling comments

Many of the modeling decisions made were for simplicity and efficiency, for instance, we have fixed d^* and m rather than allow them to be random quantities. This was done to avoid having to use a reversible jump Markov chain Monte Carlo method.

Another simplification with respect to modeling assumptions is the model we used for the covariance matrix Σ_ε of the noise, ε_{ij} , ($i, j = 1, \dots, n$). We currently model ε_{ij} as an independent random variable that is a function of the distance between two points, $d(x_i, x_j)$. A more natural approach would be to use a more sophisticated model of the covariance that would respect the fact that ε_{ij} and ε_{ik} should covary for $j \neq k$ again as a function of the distance between x_j and x_k . A full spatial model can also be proposed with

$$\Sigma_\varepsilon = \sigma_s^2 \rho(\phi_s, d_{(ij), (i'j')}) + \text{diag}(\sigma^2/w_{ij}),$$

where the first “spatial” term has a variance parameter σ_s^2 and a specified covariogram with some parameter ϕ_s and a suitable distance measure between data pairs, and the second “nugget” effect is the diagonal matrix in the model we currently use in practice. Such a model is however computationally difficult.

3.4 Simulated and real data examples

We illustrate the ideas developed and the efficacy of the method on real and simulated data. We first focus on simulated data to ground our argument. We then illustrate the utility of our approach using real data.

3.4.1 Linear regression and dimension reduction

This simple simulation based on binary linear regression model fixes the modeling ideas we have proposed with respect to dimension reduction.

The following data set was used in Mukherjee and Wu (2006). Data was generated by draws from the following two classes of samples:

$$\begin{aligned} X_{j=1,\dots,10}|y=0 &\sim N(1.5, 1), & X_{j=41,\dots,50}|y=1 &\sim N(1.5, 1), \\ X_{j=11,\dots,20}|y=0 &\sim N(-3, 1), & X_{j=51,\dots,60}|y=1 &\sim N(-3, 1), \\ X_{j=21,\dots,80}|y=0 &\sim N(0, 0.1), & X_{j=1,\dots,40,61,\dots,80}|y=1 &\sim N(0, 0.1), \end{aligned}$$

where X_j is the j -th coordinate of the 80 dimension random vector X .

Twenty samples were drawn from each class for the analysis and the data matrix is displayed in Figure 3.1(a). This is a setting where the number of variables is larger than the sample size. The posterior mean gradient outer product matrix is displayed in Figures 3.1(b). The blocking structure reflects the expected covariance with respect to the response of the predictive variables. In this example there is one dimension reduction direction and the posterior Karcher mean of this is plotted in Figure 3.1(c), in which the observation that the d.r. direction has large (positive or negative) entries in those relevant dimensions and small entries in noise dimensions is consistent with the data simulation scheme.

To illustrate mixing of the Markov chain proposed by our model we examined the

mixing of the d.r. direction from each draw from the chain. Denote $\beta_{(t)}$ as the d.r. direction for the t -th draw. We examined trace plots of these directions, $a_{(t)} = \beta'_{(t)}\beta_{(t+1)}$, where the scalar value $a_{(t)}$ is the projection of the previous eigenvector drawn onto the current direction. Figure 3.1(d) suggests that the chain is mixing and seems to be convergent.

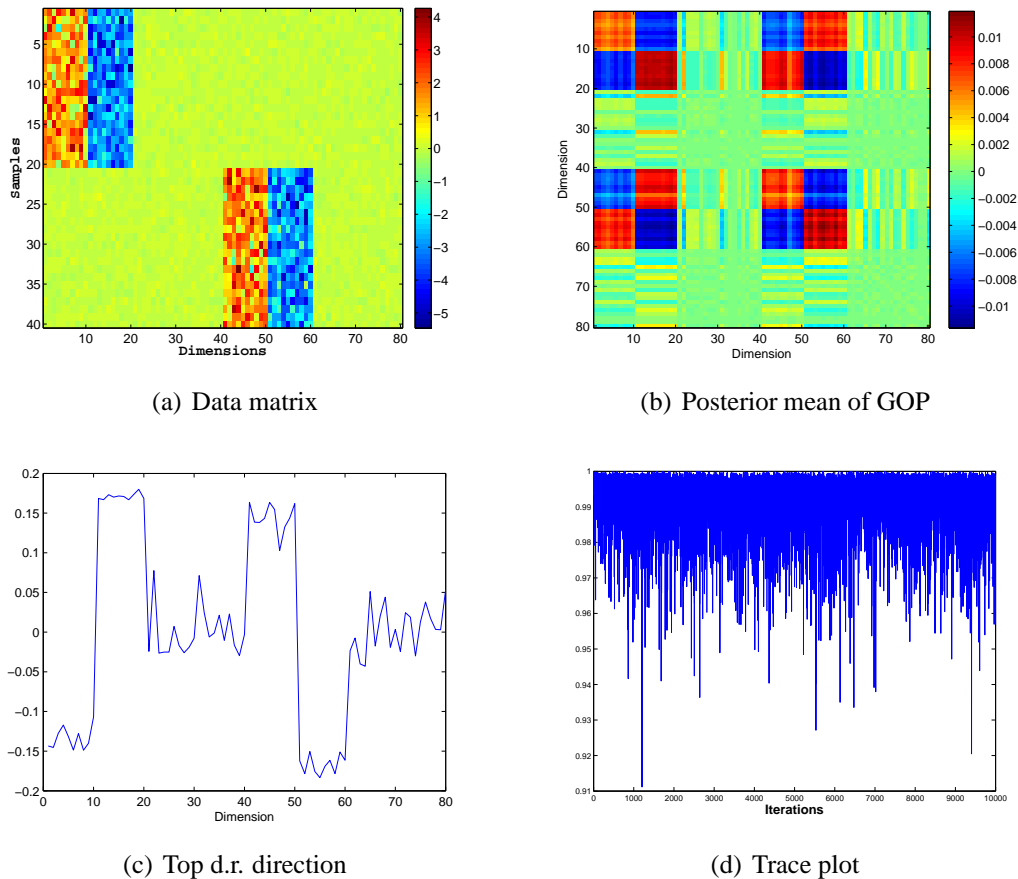


Figure 3.1: (a) The data matrix with rows corresponding to samples and columns to predictor variables (dimensions); (b) The posterior mean of the gradient outer product matrix; (c) The posterior mean of the top d.r. direction; (d) The trace plot for the inner product of two consecutive draws of the top d.r. direction.

3.4.2 Linear regression and graphical models

A simple linear regression model is used here to illustrate inference of conditional dependencies of predictor variables relevant to prediction.

The predictor variables are correspond to a five dimension random vector drawn from the following model

$$X_1 = \theta_1, X_2 = \theta_1 + \theta_2, X_3 = \theta_3 + \theta_4, X_4 = \theta_4, X_5 = \theta_5 - \theta_4,$$

where $\theta \sim N(0, 1)$. The regression model is

$$Y = X_1 + \frac{X_3 + X_5}{2} + \varepsilon,$$

where $\varepsilon \sim N(0, 0.25)$.

One hundred samples were drawn from this model and we estimated to mean and standard deviation of the gradient outer product matrix, see Figure 3.2. The posterior mean partial correlation matrix and its standard deviation are also displayed in Figure 3.2. The inference consistent with the estimate of the partial correlation structure is that X_1, X_3, X_5 are negatively correlated with respect to variation in the response and X_2 and X_4 are not correlated with respect to variation of the response. This relation is displayed in the graphical model in Figure 3.3, in which of particular interest is that we could not only depict the conditional dependence structure but also evaluate the uncertainty of our dependence estimation, as illustrated by the thickness of the edges connecting nodes.

3.4.3 Digits analysis

The MNIST digits data (<http://yann.lecun.com/exdb/mnist/>) is commonly used in the machine learning literature to compare algorithms for classification and dimension reduction. The data set consists of 60,000 images of handwritten digits 0, 1, \dots , 9 where each image is considered as a vector of $28 \times 28 = 784$ gray-scale pixel intensities. The utility of the digits data is that the effective dimension reduction directions have a visually intuitive

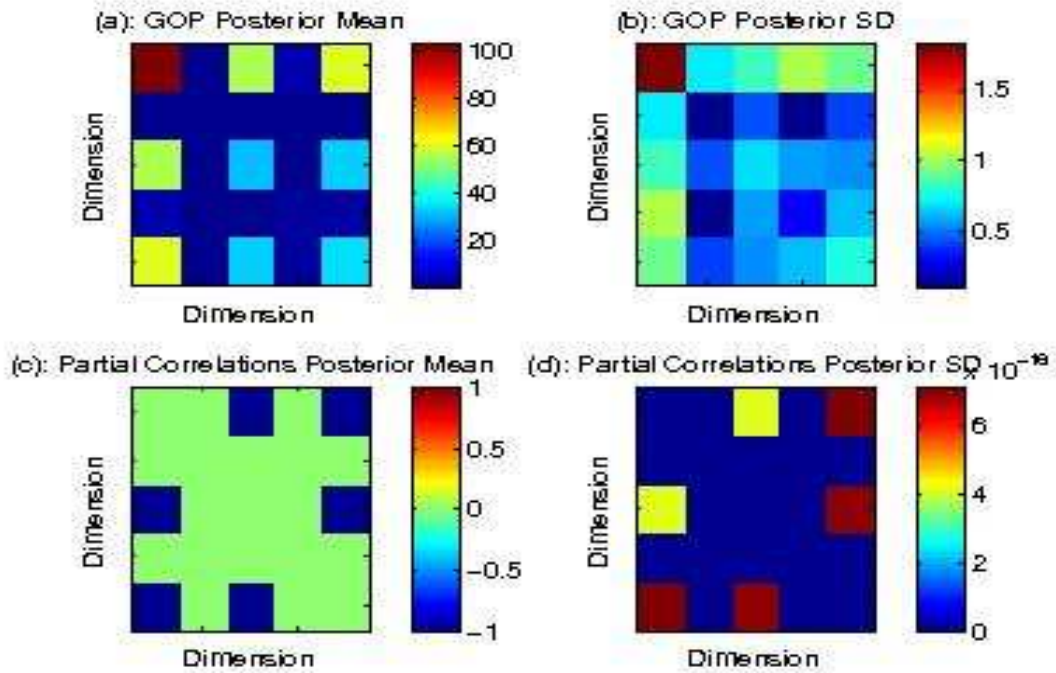


Figure 3.2: (a) and (b) are the posterior mean and standard deviation for the GOP, respectively; (c) and (d) are the posterior mean and standard deviation for the partial correlation matrix, respectively.

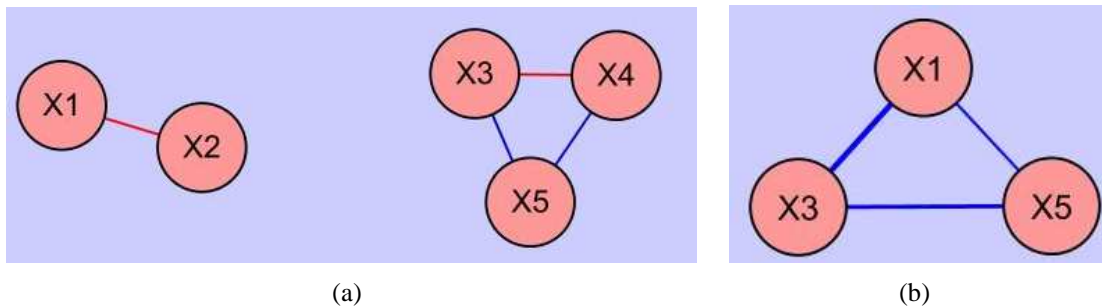
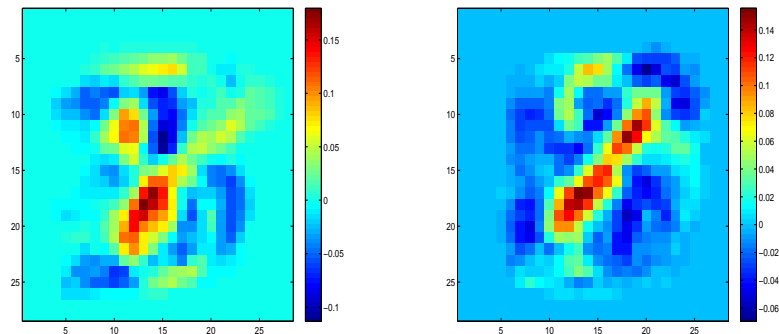


Figure 3.3: Graphical models inferred from the (a) the gradient outer product matrix and (b) the covariance matrix of the predictor variables. Each node represents a variable and each edge indicates conditional dependence. The distance of the edge is inversely proportional to the amount of dependence, the thickness of the edge is proportional to the certainty of the inference and blue edges are negative while red edges are positive.

interpretation.

For the digits data the following pairs were the most difficult to classify 3 versus 8, 2 versus 7, 4 versus 9. We examined two classification problems 3 versus 8 and 5 versus 8. For both classification problems we found that almost all of the predictive information was contained in one d.r. direction. This direction as a vector can be thought of as a 28×28 image of what is different between 3 and 8 or 5 and 8 respectively. In Figure 3.4 we display these images for 3 v.s. 8 and 5 v.s. 8, left and right panels respectively. These images were computed by applying our model to random draws of 200 samples (100 for each class) for each of the two classification problems and computing posterior Karcher mean of the top eigen-vectors of the posterior GOP matrices. 5000 posterior samples were recorded after an initial 5000 burn-in step. We see in the left panel that the upper and lower left regions are the pixels that differentiate a 3 from an 8. Similarly, for the 5 versus 8 example the diagonal from the lower left to the upper right differentiates these digits.



(a) Posterior mean of 3 vs 8

(b) Posterior mean of 5 vs 8

Figure 3.4: (a) The posterior mean of the d.r. direction for 3 versus 8, shown in a 28×28 pixel format. (b) The posterior mean of the d.r. direction for 5 versus 8, shown in a 28×28 pixel format.

3.4.4 Inference of graphical models for cancer progression

The last example is to illustrate the utility of our model in a practical problem in cancer genetics, modeling tumorigenesis. Genetic models of cancer progression are of great interest to better understand the initiation of cancer as well as the progression of disease into metastatic states. In Edelman, Guinney, Chi, Febbo, and Mukherjee (Edelman et al.) a models of tumor progression in prostate cancer as well as melanoma were developed. One fundamental idea here was that the predictor variables were summary statistics that assayed the differential enrichment of a priori defined sets of genes in individual samples (Edelman et al., 2006; Edelman, Guinney, Chi, Febbo, and Mukherjee, Edelman et al.). These a priori defined gene sets correspond to genes known to be in signalling pathways or have functional relations. The other fundamental idea is an inference of the interaction between pathways as the disease progresses across stages.

A variation of the analysis in Edelman, Guinney, Chi, Febbo, and Mukherjee (Edelman et al.) is developed in this section. The data consists of 22 benign prostate samples and 32 malignant prostate samples. For each sample we compute the enrichment with respect to 522 candidate gene sets or pathways (Mukherjee and Wu, 2006). Each sample corresponds to a 522 dimensional vector of pathway enrichment. We applied our model to this data set and inferred a mean posterior conditional independence matrix as well as the uncertainty in the estimates of these elements. For visualization purposes we focused on the 16 pathways most relevant with respect to predicting progression. For these pathways we plot the conditional independence matrix and the variance of the elements in the matrix in Figure 3.5. Red edges correspond to positive partial correlations and blue for negative. The width of the edges correspond to the degree of uncertainty: edges we are more sure of are thicker. This graph offers a great deal of interesting biology to explore some of which is known, see Edelman, Guinney, Chi, Febbo, and Mukherjee (Edelman et al.) for more details. One

particularly interesting aspect of the inferred network is the interaction between the ErbB (ERBB) or epidermal growth factor signalling pathway and the mitogen-activated protein kinase (MAPC) pathway.

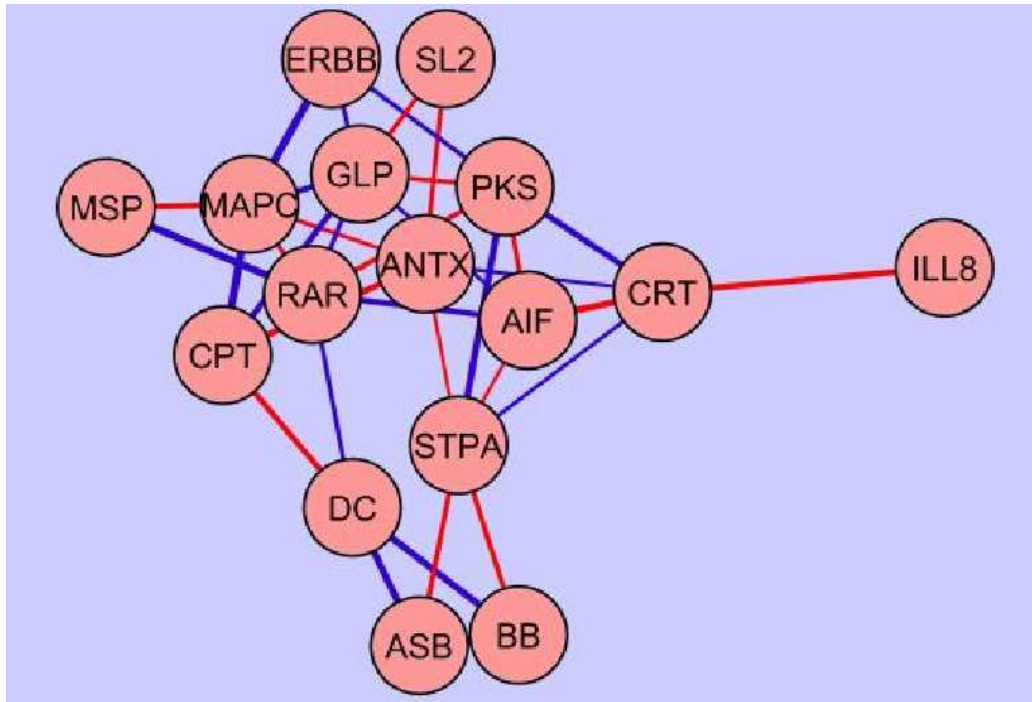


Figure 3.5: The association graph for the progression of prostate cancer from benign to malignant based on the inferred partial correlation. Red edges correspond to positive partial correlations and blue for negative. The width of the edges correspond to the degree of uncertainty, edges we are more sure of are thicker.

3.5 Discussion

We propose a Bayesian nonparametric model for simultaneous dimension reduction and regression as well as inference of graphical models. This approach applies to the classical Euclidean setting as well as nonlinear manifolds. We illustrate the efficacy and utility of this model on real and simulated data. An implication of our model is that there are fascinating connections between spatial statistics and manifold and nonlinear dimension

reduction methods that should be of great interest to the statistics community.

Chapter 4

Nonparametric Bayesian Kernel Models for Regression and Classification

4.1 Introduction

In regression problems the main purpose is to infer the relationship between a response variable $Y \in \mathcal{Y} \subset \mathbb{R}$ and predictors or explanatory variables $X \in \mathcal{X} \subset \mathbb{R}^p$. The regression model is typically summarized by $Y = f(X) + \varepsilon$ by additive noise assumption with f an unknown regression function and ε some noise, and the goal is to infer through (training) data consisting of n observations $y_i \in \mathbb{R}, x_i \in \mathbb{R}^p (i = 1, \dots, n)$ the regression function f which can further be used for future prediction given new predictor variable values. Nonparametric models, unlike parametric models that directly specify a parametric form for f , generally put f into a suitable framework and learn f therein. An important class is the kernel models which have been used extensively in machine learning for classification and regression problems (Hastie et al., 2001; Schölkopf and Smola, 2001), with some well known examples including spline models (Wahba, 1990) and support vector machines (SVMs) (Cortes and Vapnik, 1995). The appealing properties of kernel models are their flexibility and predictive accuracy, and most importantly, their ability to handle high dimensional data.

As formulated in Section 1.2 nonparametric kernel models proceed by constructing a so called reproducing kernel Hilbert space (RKHS) \mathcal{H}_k (Wahba, 1990) generated by a positive semi-definite kernel function $k(x, u)$ with $x, u \in \mathcal{X}$, the input space, and selecting

the estimate in \mathcal{H}_k . Regularization methods are frequently used to justify the estimate as in (1.4), leading to the finite dimensional representation form (1.5) due to the representer theorem (Kimeldorf and Wahba, 1971). This reduces an infinite dimensional optimization problem to one in n variables, which is very attractive for high-dimensional analysis since the optimization is over $n \ll p$ variables and independent of the dimension p .

Access to fully Bayesian formulations of kernel methods would provide a natural framework to further the richness and interpretability of kernel models – a program driving much research in data mining and machine learning. A Bayesian approach adds in probabilistic interpretations and provides rich inference such as easy evaluation of uncertainty for estimates and prediction through posterior samples and would allow for the immediate relaxation of the limitation for the above penalized loss functional framework of requiring additional methods such as bootstrapping or cross-validation to provide confidence intervals and set hyper-parameters. Bayesian kernel methods, in particular Bayesian SVMs, have certainly been proposed (Tipping, 2001; Sollich, 2002) based on applying Bayesian estimation directly to the finite representation from equation (1.5). However, the direct adoption of (1.5) is not a proper statistical model, as the model changes with the sample size and observed covariate values defining the knots, without a coherent argument. One may justify the use of (1.5) in a Bayesian analysis by the fact that the finite representation is a MAP estimator (Wahba, 1990) for certain priors over the entire RKHS, though priors used in these referenced Bayesian analysis are not the same as the priors used to establish the connection between equation (1.5) the MAP estimator.

Our goal here is to provide a novel, fully Bayesian framework and theory for kernel regression and classification that address the above issues. Unlike previous approaches we specify priors on the entire RKHS. Our prior specification induces a class of functions that span the RKHS, providing an equivalence between the nonparametric Bayesian model and kernel models used in the penalized loss framework. This implies a Bayesian representer

theorem that results in the finite representation in equation (1.5) derived from a Bayesian formulation, and that is coherent across samples and sample sizes. This formal model then easily and coherently addresses problems of inference on hyper-parameters, variable selection, and ancillary issues such as unlabeled data (in semi-supervised learning).

Section 4.2 describes the nonparametric Bayesian approach that allows us to place a coherent prior on the RKHS and recover the parametrisation of the representer theorem as an approximation of the posterior mean. Section 4.3 provides one approach to complete prior specification over model hyper-parameters and a corresponding MCMC approach to posterior evaluation and inference for both regression and classification settings. Section 4.4 extends the definition of kernels to allow for variable (or “feature”) selection in the covariate space. Examples and discussion are given in Section 4.5, with summary comments in Section 4.6.

4.2 A Class of Non-Parametric Bayesian Kernel Models

In this section we develop kernel models that are based on integral operators placing priors on signed measures rather than directly on the regression function space. We first show why we do not elicit priors directly on the function space.

4.2.1 Problems with Direct Prior Elicitation

There are two natural ways to directly elicit priors on a RKHS, \mathcal{H}_K , induced by a kernel k : based on orthogonal expansions of the RKHS, or based on the duality between Gaussian processes and RKHS. Both have drawbacks.

Kernel functions $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that are continuous and positive semi-definite on a compact space \mathcal{X} are Mercer kernels for which the RKHS is characterized (Mercer, 1909)

as

$$\mathcal{H}_K = \left\{ f \mid f(x) = \sum_{j=1}^{\infty} a_j \phi_j(x) \text{ such that } \sum_{j=1}^{\infty} a_j^2 / \lambda_j < \infty \right\},$$

where $\{\lambda_j\}$ and $\{\phi_j(x)\}$ are the eigenvalues and eigenfunctions of the integral operator defined by the kernel function

$$\lambda_j \phi_j(x) = \int_{\mathcal{X}} k(x, u) \phi_j(u) d\mu(u),$$

where μ is a measure. The eigenvalues and RKHS do not depend on the measure, so a prior over the space $\mathcal{A} = \{(a_j)_{j=1}^{\infty} \mid \sum_{j=1}^{\infty} a_j^2 / \lambda_j < \infty\}$ implies a prior on \mathcal{H}_K . There are serious computational and conceptual problems with specifying a prior on the parameter \mathcal{A} ; it is in general infinite-dimensional, and it is subject to challenging constraints. The crux of the problem is that in this orthonormal expansion model we are working explicitly with eigenfunctions and eigenvalues, and they are inherently challenging to manipulate; many popular kernels do not even lead to eigenfunctions with closed forms, and others are not even computable.

The Gaussian process view exploits the duality between RKHS and Gaussian processes (Wahba, 1990) to suggest a natural prior specification that has been used extensively in Bayesian modelling. A prior can be placed directly on a space of functions by sampling from the paths of the Gaussian process with covariance structure defined by k . The problem with this approach is that the random functions drawn are almost surely outside the RKHS induced by k (Wahba, 1990), so that using this approach in fact does not characterize \mathcal{H}_k .

4.2.2 Priors and Integral Operators

Alternatively, consider the space of functions defined as a convolution of the kernel with a signed (Borel) measure

$$\mathcal{G} = \left\{ f \mid f(x) = \int k(x, u) d\gamma(u), \gamma \in \Gamma \right\}, \quad (4.1)$$

with $\Gamma(\cdot)$ as a subset of the space of signed Borel measures. Equivalence between \mathcal{G} and \mathcal{H}_k exists for appropriate choices of priors on Γ (Liao, 2005; Pillai et al., 2007) including Dirichlet Process Priors (Ferguson, 1973, 1974), so that in order to elicit priors on \mathcal{H}_k we now need priors on \mathcal{G} . Placing a prior on Γ implies a prior on \mathcal{G} .

A variation of the integral operator defined in (4.1) takes the form

$$f(x) = \int k(x, u) d\gamma(u) = \int k(x, u) w(u) dF(u), \quad (4.2)$$

where the random signed measure $\gamma(\cdot)$ is decomposed into a probability distribution $F(\cdot)$ and coefficient function $w(\cdot)$; F and γ share the same support. We now need priors for both w and F . In general F denotes the distribution of the location of kernel knots u . Here we set $F = F_X$, the marginal distribution of X . This is a reasonable assumption as long as F_X and γ share the same support. An appealing property of this dependence of f on F_X is that our estimate of $f(x)$ will be locally adaptive in that more knots are allocated in high density regions.

4.2.3 Dirichlet Process Priors

The Dirichlet process (DP) prior is a natural choice to model uncertainty about the distribution function F . For a specified distribution F_0 having the same support as the uncertain distribution F , and a positive scale parameter α , the notation $\text{DP}(\alpha, F_0)$ implies

that for any measurable partition of the sample space (B_1, B_2, \dots, B_k) , the random vector $(F(B_1), \dots, F(B_k))$ follows a Dirichlet distribution with parameter $\alpha(F_0(B_1), \dots, F_0(B_k))$ (Ferguson, 1973, 1974; Sethuraman, 1994). DP priors are very popular in practical non-parametric Bayesian analysis (Escobar and West, 1995; MacEachern and Müller, 1998) due to modeling flexibility and computational advantages.

A fundamental characteristic of the DP model is that, given a sample $X_n = (x_1, \dots, x_n)$ drawn independently from (uncertain) distribution F , the posterior is also DP

$$F | X_n \sim \text{DP}(\alpha + n, F_n), \quad F_n = (\alpha F_0 + \sum_{i=1}^n \delta_{x_i}) / (\alpha + n). \quad (4.3)$$

Consider, then, such a prior for F in equation (4.2), and choose some fixed new point x_* to predict the function value $f(x_*)$. Based on the sample of n draws X_n from F we see that

$$\mathbb{E}[f | X_n] = a_n \int k(x, u) w(u) dF_0(u) + n^{-1}(1 - a_n) \sum_{i=1}^n w(x_i) k(x, x_i)$$

where $a_n = \alpha / (\alpha + n)$. Taking the limit of $\alpha \rightarrow 0$ to represent a non-informative prior leads to the finite-dimensional *Bayesian Representer* form

$$\hat{f}_n(x) = \sum_{i=1}^n w_i k(x, x_i), \quad (4.4)$$

where $w_i = w(x_i)/n$ depends on the “knot” x_i and sample size n . The two finite representations, equations (1.5) and (4.4), take the same form although they are derived from two fundamentally different approaches: the solution of a Tikhonov regularization functional versus formal process-prior Bayesian modeling.

4.3 Estimation and Inference

4.3.1 Likelihood and Prior Specification for Hyper-Parameters

The Bayesian representor form leads to the usual linear regression on the kernel values as predictors with regression parameters w_i . Adding an intercept and a normal error model assumption we have the standard form

$$y_i = w_0 + f(x_i) + \varepsilon_i = w_0 + \sum_{j=1}^n w_j k(x_i, x_j) + \varepsilon_i, \quad (i = 1, \dots, n), \quad (4.5)$$

where $\varepsilon_i \sim N(0, \sigma^2)$. In vector form, the model is

$$Y \sim N(w_0 \mathbf{1} + Kw, \sigma^2 I) \quad (4.6)$$

where $\mathbf{1} = (1, \dots, 1)'$, K is the $n \times n$ design matrix having entries $k(x_i, x_j)$, $Y = (y_1, \dots, y_n)'$ and the regression parameter vector is $w = (w_1, \dots, w_n)'$.

Traditional priors can be taken for (w_0, σ^2) and we use the standard reference prior component $\pi(w_0, \sigma^2) \propto 1/\sigma^2$. Specifying priors over the w_i can be done by defining sample size independent priors for values $w(x_i)$ at arbitrary knots and then scaling by $1/n$. As an alternative, we can induce appropriate sample size dependence and address key questions of inducing regression shrinkage appropriately coupled to the structure of the kernel design space by using ridge regression or g-prior modeling (Zellner, 1986). West (2003) defined and exemplified the use of a flexible and practically very effective class of generalised g-priors that allow for different degrees of shrinkage estimation of regression parameters in different principal component directions on the induced design space for any regression model, and we adopt that strategy here. This is particularly relevant when dealing with many predictors, as it provides an ability to “shrink away” the effects of many irrelevant

component dimensions while highlighting those of predictive value. This class of priors explicitly models the distribution $p(w|K)$, so that the sample size dependence is directly induced and the class of priors adapts as the sample size changes.

Specifically, a generalised g-prior is induced by independent normal priors on the regression parameters of the equivalent principal component regression transformation of the model. The kernel matrix K is symmetric and positive semi-definite, so has spectral decomposition $K = \tilde{F}\Delta\tilde{F}'$ where \tilde{F} is the $n \times n$ orthogonal *factor* matrix, and $\Delta = \text{diag}(\lambda_1^2, \dots, \lambda_n^2)$. In the orthogonal representation the regression maps from Kw to $F\beta$ with $F = \tilde{F}\Delta$, $w = \tilde{F}\beta = F\Delta^{-1}\beta$. Assume conditionally independent normal priors for the elements of β , so that $\beta \sim N(0, T)$ for some $T = \text{diag}(\tau_1, \dots, \tau_n)$. The induced generalised g-prior for w is then

$$(w | K, T) \sim N(0, F\Delta^{-1}T\Delta^{-1}F'). \quad (4.7)$$

Following West (2003), we further specify hyper-priors over the n prior variances τ_i – that play roles as shrinkage parameters – as independent inverse gammas,

$$\tau_i^{-1} \sim \text{Gamma}(a_\tau/2, b_\tau/2), \quad (i = 1, \dots, n),$$

inducing heavier-tailed t-priors on the w_i when we marginalize over the τ_i .

Viewed as hyper-parameters to be estimated, the τ_j 's are the prior variances for each factor regression parameter and allow for a varying degree of shrinkage in each of the orthogonal factor dimensions, as discussed. Hyper-parameter values $a_\tau = b_\tau = 2$ correspond to Cauchy distributions on the β_j , a very natural, highly diffuse though proper prior specification. Note that in practice, for computational stability, we may choose to truncate the spectral decomposition by rejecting factors with very small eigenvalues λ_i ; that is, in such

a case we may choose to replace \tilde{F} (hence also F) by its first m columns, Δ by its first m diagonals, and set $T = \text{diag}(\tau_1, \dots, \tau_m)$, where $m < n$ and the eigenvalues $\lambda_{m+1}, \dots, \lambda_n$ are less than some pre-specified threshold.

As a practical modification, the positivity of kernels suggests that we center the induced kernel predictors for both interpretation and numerical stability, in advance of developing the above analysis. That is, $k(\cdot, \cdot)$ is replaced by a centred kernel $\tilde{k}(\cdot, \cdot)$ with

$$\tilde{k}(x_i, x_j) = k(x_i, x_j) - \bar{k}_{i.} - \bar{k}_{.j} + \bar{k},$$

where $\bar{k} = \sum_{i,j=1}^n k(x_i, x_j)/n^2$, $\bar{k}_{i.} = \sum_{j=1}^n k(x_i, x_j)/n$ and $\bar{k}_{.j} = \sum_{i=1}^n k(x_i, x_j)/n$. The kernel covariate matrix K is then redefined with entries $K_{ij} = \tilde{k}(x_i, x_j)$. It is evident that this is positive semi-definite, and the singular value decomposition and the ensuing analysis as detailed above proceed otherwise unchanged.

4.3.2 Model Fitting and Prediction via MCMC

Given the likelihood and the prior distributions a standard Gibbs sampler can be used to simulate the posterior $p(w_0, w, \sigma^2 \mid \text{data})$. After initialization, samples of parameters and hyper-parameters are drawn sequentially from the complete conditional posterior distributions. At each iteration, with all relevant conditioning parameters fixed at their most recent values in the iterates, we update as follows:

1. Update w_0 : w_0 is drawn from the normal posterior with mean $n^{-1}\mathbf{1}'(Y - F\beta)$ and variance σ^2/n .
2. Update w : Simply via β , generate $\beta \sim N(b, V)$ where $V = \text{diag}(V_1, \dots, V_m)$ with $V_i = \sigma^2\tau_i/(\tau_i + \sigma^2)$, and $b = VF'(Y - w_0)/\sigma^2$; then set $w = F\Delta^{-1}\beta$.
3. Update T : For $j = 1, \dots, m$, $\tau_j^{-1} \sim \text{Gamma}((a_\tau + 1)/2, (b_\tau + \beta_j^2)/2)$.

4. Update σ^2 : $\sigma^{-2} \sim \text{Gamma}(n/2, s/2)$ with $s = e'e$ where $e = Y - w_0 - F\beta$.

For prediction at a specified new point x_* , any aspect of the predictive distribution for y_* can be included in the MCMC sampling. Suppose the sampled parameter values are $(w_0^{(t)}, w^{(t)}, \sigma^{2(t)})$ at each iteration $t, t = 1, \dots, T_c$, the MC approximation to the predictive distribution is simply

$$\begin{aligned} p(y_*|x_*, \text{data}) &= \int p(y_*|x_*, w_0, w, \sigma^2)p(w_0, w, \sigma^2|\text{data})d(w_0, w, \sigma^2) \\ &\approx \frac{1}{T_c} \sum_{t=1}^{T_c} p(y_*|x_*, w_0^{(t)}, w^{(t)}, \sigma^{2(t)}) \end{aligned}$$

We can also evaluate the mean of the conditional normal distribution $p(y_*|x_*, w_0^{(t)}, w^{(t)}, \sigma^{2(t)})$ to compute MC approximations to predictive mean $\mathbb{E}(y_*|x_*, \text{data})$. We can do this across a range of x_* values to map out the predicted non-linear regression function for predictive uses.

4.3.3 Binary Regression for Classification

The approach developed above for regression models can of course be easily extended to a classification setting using probit regression, or other binary regressions. The standard latent variable imputation extensions of MCMC lead directly to posterior samplers for probit link function. Metropolis-Hastings variants for logistic regression are also trivial modifications. These are practically very relevant extensions for kernel classification problems.

By way of notation and basic structure in the probit model, the responses y_i in the kernel model (4.6) are now latent and the normal errors are standard, i.e., $\sigma^2 = 1$. We observe binary responses $Z = (z_1, z_2, \dots, z_n)'$ generated by $z_i = 1(0)$ if $y_i \geq 0 (< 0)$. The MCMC extensions simply include the latent Y values at each iteration of the simulation.

The traditional Gibbs sampler iterates between sampling conditional posteriors for Y given the regression parameters, and vice-versa. Though often effective, this vanilla Gibbs sampler can suffer from very slow mixing due to high correlations between successive draws of latent variables. We will develop a novel and effective solution: rather than the Gibbs sampler we use a Metropolis-Hastings method that samples the kernel model parameters jointly with the latent variable Y . This is summarised in the following section, in an extended model that incorporates additional kernel parameters to address variable and feature selection.

4.4 Variable and Feature Selection

4.4.1 Kernel Model Extension

Variable selection and feature selection are important problems in high-dimensional regression. The standard formulation of variable selection is to select a relatively small subset of the p predictors without loss of predictive accuracy. In the problem of feature selection a small subset of combinations of the p predictors are selected. Principle components regression with a few principle components is an example of a feature selection method.

Standard practise in kernel regression allows each coordinate of x to be scaled (Chapelle et al., 2002), i.e.,

$$k_\rho(x, u) = k(\sqrt{\rho} \otimes x, \sqrt{\rho} \otimes u)$$

where $a \otimes b$ is the element-wise product of two vectors and $\rho = (\rho_1, \dots, \rho_p)$ is a p -dimensional vector with $\rho_k \in [0, \infty]$ as an individual scale parameter for the k -th dimension. This approach can be applied to most kernels and for the linear, polynomial, and

Gaussian kernels the resulting adaptive kernels are

$$\begin{aligned}
 k_\rho(x, u) &= \sum_{k=1}^p \rho_k x_k u_k, \\
 k_\rho(x, u) &= \left(1 + \sum_{k=1}^p \rho_k x_k u_k \right)^d, \\
 k_\rho(x, u) &= \exp \left(- \sum_{k=1}^p \rho_k (x_k - u_k)^2 \right).
 \end{aligned}$$

We will focus on the Gaussian kernel for which ρ_k can be regarded as the reciprocal of the bandwidth for the k -th variable which determines the neighborhood size for that dimension. When $\rho_k = 0$, the neighborhood size is infinity and the corresponding variable is irrelevant in predicting the response variable. Variable/bandwidth selection is then a problem of estimation or selection of the parameter ρ , now explicitly in the context of allowing for zero values. This invites analysis using standard Bayesian variable selection/model uncertainty strategies based on “point mass, mixture prior” over these parameters.

For each ρ_k independently, we adopt the prior

$$\begin{aligned}
 \rho_k &\sim (1 - \gamma)\delta_0 + \gamma \text{Gamma}(a_\rho, a_\rho s), \quad (k = 1, \dots, p), \\
 s &\sim \text{Exp}(a_s), \quad \gamma \sim \text{Beta}(a_\gamma, b_\gamma)
 \end{aligned}$$

where $(a_\rho, a_s, a_\gamma, b_\gamma)$ are specified hyperparameters, $\text{Beta}(\cdot, \cdot)$ represents the beta distribution and $\text{Exp}(\cdot)$ the exponential.

4.4.2 Overall MCMC

The MCMC analysis can now be extended to include the ρ parameters. These parameters are treated with appropriate Metropolis-Hastings steps since their complete conditionals are not of standard forms. Our overall MCMC sampler uses a Metropolis-Hastings

step to jointly sample the kernel bandwidth and regression parameters; in the case of binary outcomes when the y_i are latent, this sampling step is extended to jointly sample these parameters and Y together. As mentioned in the previous section, this novel MCMC – that has been tested successfully in a number of examples – is designed to mix faster than the traditional Gibbs sampler in binary models, and now also provides an overall approach for the kernel variable selection extension in both continuous and binary outcomes cases.

The full hybrid sampler for the posterior of $(w_0, \beta, \rho, Y, T, s, \gamma)$ in the case of binary probit regression is detailed here. The changes to this to generate the corresponding MCMC for the continuous-response regression case simply adds in the sampling of the residual variance σ^2 at each step and removes the imputation of the y_i that are known.

The sequence of steps per iteration in the full binary kernel model with feature selection are as follows:

1. Update w_0 as in section 4.3.2.
2. Update (ρ, β, Y) jointly, in the following two steps.
 - 2.1. Update (ρ, β) :
 - 2.1.1. Propose ρ^* : Let p_g, p_l, p_u denote the probabilities for a *global move*, *local move*, or *update move* respectively.
 - For the *global move*, draw ρ^* from the prior.
 - For the *local move*, set $\rho^* = \rho$ then randomly pick a dimension k . If $\rho_k \neq 0$, set $\rho_k^* = 0$; otherwise draw $\rho_k^* \sim \text{Gamma}(a_\rho, a_\rho s)$, the continuous part of the prior.
 - For the *update move*, set $\rho^* = \rho$ and then, for all dimensions k where $\rho_k \neq 0$, draw $\rho_k^* \sim \text{Gamma}(a_\rho, a_\rho s)$.

Our proposals use $p_g = .25, p_l = .5, p_u = .25$.

2.1.2. Propose β^* : Compute the proposed kernel matrix K^* with entries $\tilde{k}_{\rho^*}(x_i, x_j)$ and its spectral factors F^* and Δ^* . Set $\hat{Y} = w_0 + F^*\beta$ and simulate Y^* via, for each $i = 1, \dots, n$,

$$y_i^* \sim \begin{cases} N(\hat{y}_i, 1)^+, & \text{if } z_i = 1, \\ N(\hat{y}_i, 1)^-, & \text{if } z_i = 0. \end{cases}$$

Then, propose $\beta^* \sim N(b^*, V)$ where $V = \text{diag}(V_1, \dots, V_m)$ with $V_i = \tau_i/(1 + \tau_i)$, and $b^* = VF^{*'}(Y^* - w_0)$.

2.1.2. Acceptance ratio to compare and test (ρ^*, β^*) against the current values (ρ, β) : The Metropolis-Hastings acceptance ratio is

$$r = \frac{p(Z | \rho^*, \beta^*, w_0) \pi(\rho^*, \beta^* | s, \gamma, T) q(\rho, \beta | T, w_0, s, \gamma)}{p(Z | \rho, \beta, w_0) \pi(\rho, \beta | s, \gamma, T) q(\rho^*, \beta^* | T, w_0, s, \gamma)}$$

where the terms $p(Z | \dots)$ are likelihood evaluations from the binary regression

$$p(Z | \rho^*, \beta^*, w_0) = \prod_{i=1}^n \Phi(\hat{y}_i)^{z_i} (1 - \Phi(\hat{y}_i))^{1-z_i},$$

with $\Phi(\cdot)$ being the standard normal c.d.f, and $\pi(\cdot), q(\cdot)$ denote the prior distribution function and proposal distribution function, respectively, with

$$\pi(\rho^*, \beta^* | \gamma, s, T) = \prod_{k=1}^p \pi(\rho_k^* | \gamma, s) \prod_{j=1}^m \pi(\beta_j^* | \tau_j).$$

With probability $\min\{r, 1\}$ accept the proposed values and hence set $\rho = \rho^*$ and $\beta = \beta^*$; otherwise, retain the current values.

Denote the accepted or retained values by $\{\rho, \beta, K, F, \Delta\}$, and set $w = F\Delta^{-1}\beta$.

2.2. Update Y : $\hat{Y} = w_0 + F_\rho \beta$ and resample Y via, for each $i = 1, \dots, n$,

$$y_i \sim \begin{cases} N^+(\hat{y}_i, 1), & \text{if } z_i = 1, \\ N^-(\hat{y}_i, 1), & \text{if } z_i = 0. \end{cases}$$

where N^+ and N^- denote the positive and negative parts of a truncated normal.

3. Update hyper-parameters: (s, γ, T)

3.1. Update s : $s \sim \text{Gamma}(a_\rho + 1, a_s + a_\rho \sum \rho_k)$.

3.2 Update γ : $\gamma \sim \text{Beta}(a_\gamma + p_1, b_\gamma + p - p_1)$ where p_1 is the number of nonzero elements in ρ .

3.3 Update T as in section 4.3.2.

This MCMC combines variable and feature selection. It is a variable selection method since only those variables with nonzero ρ_k will be selected. It is also a feature selection method since we are weighting each variable that is selected by ρ_k . The parameter $m \leq n$ was introduced in Section 4.3.1 to allow for the opportunity to truncate the expansion of the kernel matrix for numerical stability. Reducing m may be problematic since principal components of the kernel matrix that dominate variation in the kernel design space may not necessarily correspond to the factors most relevant in prediction of the response variable. In the current setting that now includes the point-mass mixture priors over elements of ρ , this problem is obviated. It is still generally desirable to consider restricting to $m < n$ for numerical and computational reasons.

4.5 Examples

The following experiments on simulated and real data provide for our method an indication of the efficacy in variable selection and the accuracy reflecting the uncertainty in the estimates. For convenience our method will be referred to as “BAKER”, acronym for “BAYesian KERnels”.

4.5.1 Simulated nonlinear classification – variable selection

A data set was generated to illustrate the variable selection aspect of BAKER. The data is perfectly separable by a nonlinear classifier and has two relevant dimensions and ten noise dimensions.

Samples from class 0 were drawn from

$$(x^1, x^2) = (r \sin(\theta), r \cos(\theta)), \text{ where } r \sim \text{Unif}[0, 1] \text{ and } \theta \sim \text{Unif}[0, 2\pi],$$
$$x^j \sim \text{Unif}[-1.5, 1.5], \text{ for } j = 3, \dots, 12.$$

Samples from class 1 were drawn from

$$(x^1, x^2) = (r \sin(\theta), r \cos(\theta)), \text{ where } r \sim \text{Unif}[1, 2] \text{ and } \theta \sim \text{Unif}[0, 2\pi],$$
$$x^j \sim \text{Unif}[-1.5, 1.5], \text{ for } j = 3, \dots, 12.$$

A draw of the first two dimensions of the data is shown in Figure (4.1 a). The first two dimensions are signal with class 1 contained in the unit circle and class 0 contained in an annulus outside the unit circle.

We drew 60 training and 60 test samples, half for each class. A Gaussian kernel was

used and hyper-parameters were set to be:

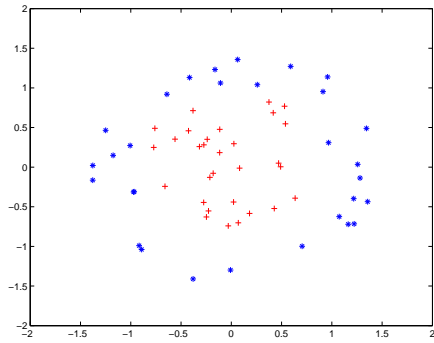
$$a_\tau = b_\tau = 2, a_\gamma = b_\gamma = 3, a_\rho = 1, a_s = 0.5;$$

These hyper-parameters are generally insensitive with respect to prediction. We ran 2000 iterations with the first 1000 burned in. The empirical training and test errors were 0% and 3%, respectively. The posterior mean of the reciprocal bandwidth parameter ρ is shown in Figure (4.1 b). That ρ_1 and ρ_2 clearly dominate those for the other noise dimensions demonstrates the efficacy for BAKER in the variable selection aspect. The prediction on the region $[-2, 2] \times [-2, 2]$ for the first two dimensions is shown in Figure (4.1 c): The boundary is approximately the unit circle, the optimal separating boundary. For a comparison Figure (4.1 d) shows the prediction on this region *without* variable selection (section 4.3.2) which fails completely.

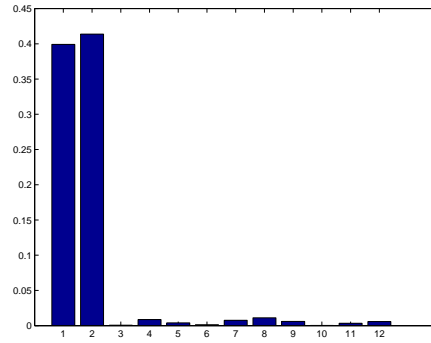
4.5.2 High-dimensional gene expression data – uncertainty in predictions

We consider a high-dimensional gene expression data set. The purpose of this analysis is to highlight the utility of BAKER for addressing the uncertainty in prediction, and also to compare BAKER with an extensively used classifier, the SVM. In this example we do not include variable selection procedure.

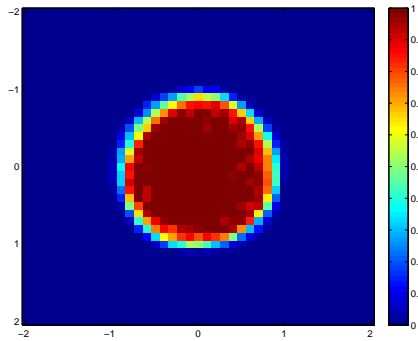
The data set Ramaswamy et al. (2001); Mukherjee et al. (2003) consists of 280 samples from patients of which 190 are tumor samples and 90 are normal. For each sample expression data from microarray analysis for 16063 genes was collected. The data was randomly split into training and test sets with 180 and 100 samples respectively. For BAKER hyper-parameters are $a_\tau = b_\tau = 2$ and 2000 iterations were run with the first 1000 burned in. A linear kernel was used for both BAKER and SVM with accuracy over twenty random



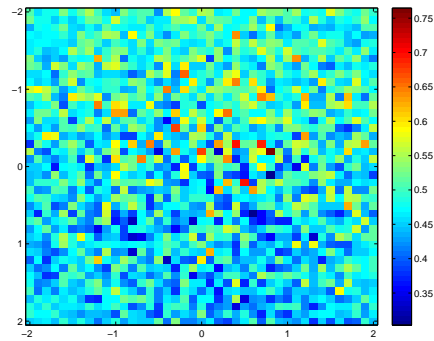
(a) Data



(b) Posterior mean of ρ



(c) Posterior prediction with variable selection



(d) Posterior prediction without variable selection

Figure 4.1: (a) Illustration for the first two dimensions of the data. Points from class 1 are red “pluses” and points from class 0 are blue “stars”. (b) The posterior means for ρ : the first two signal dimensions have large values. (c) The posterior prediction probability *with* variable selection over all points in the first two dimensions, $[-2, 2] \times [2, 2]$. (d) The same as (c) but *without* variable selection.

test-train splits reported in Table 4.1.

Table 4.1: The training and test accuracies for different methods on the gene expression data

	Training Accuracy (standard deviation)	Test Accuracy (standard deviation)
BAKER	97.2% (1.0%)	88.3% (4.1%)
SVM	98.9% (0.5%)	91.3% (2.8%)

In terms of point estimation, they are comparable while SVM shows certain superiority. The true utility of BAKER is that it provides a predictive distribution and not just a point estimate as SVM would do. We illustrate this on one random test-train split where the test error was 13%. In Figure (4.2) we plot the mean of the posterior distribution for each test sample in red stars and provide a point-wise 90% credible interval for each test sample. A result of this analysis is that in addition to misclassified instances there are 7 correctly classified samples with great uncertainty (credible intervals covering 0.5) that may require further investigation.

4.5.3 UCI Data sets

We further examine variable selection and the predictive posterior distribution on several data sets from the UCI Machine Learning Repository¹.

Wisconsin Breast Cancer Dataset

The Wisconsin Breast Cancer data was examined in Zhang (1992) using a variety of classifiers. The class labels were benign and malignant and after removal of samples with missing attributes 353 samples remained. The input data had nine attributes. The best result reported in Zhang (1992) with a training set of 200 samples with 100 from each class and

¹<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

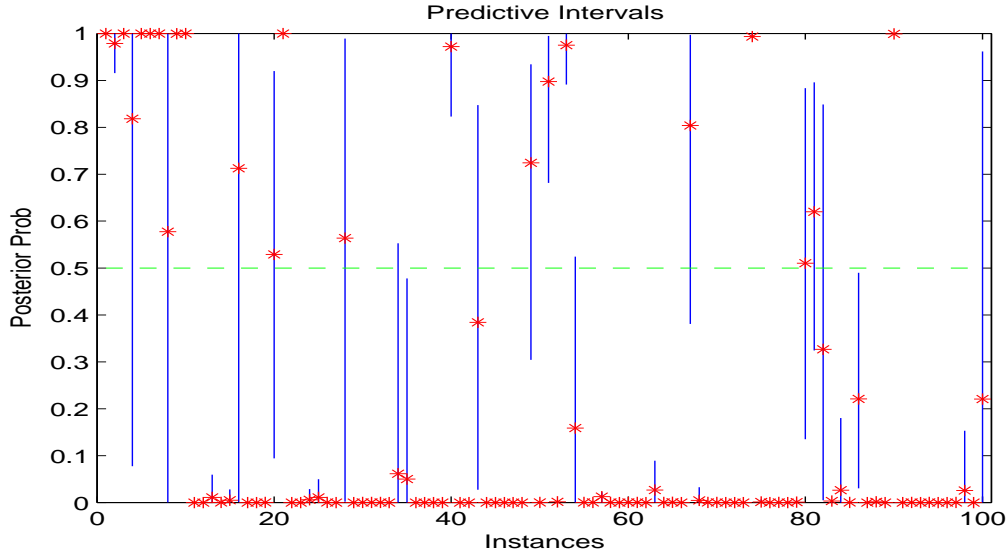


Figure 4.2: The posterior predictive distribution for a test set with the first 10 samples are normal and the remaining tumor. The red stars represent the posterior means and the blue lines are 90% credible intervals. There are 13 cases that are misclassified and 7 more that are very uncertain.

the remaining samples as test was 93.7% by 1 nearest neighbor. We repeated this comparison using BAKER (specifications of hyper-parameters follow those in section(4.5.1); 1000 iterations with the initial 500 burned in), SVM, and a Generalized Linear Model (GLM) with a probit link as the classification algorithms over ten random 200-153 training-test splits. The classifiers perform comparably as in Table 4.2. In terms of point estimation they are all comparable.

Table 4.2: The training and test accuracies for different methods on the Wisconsin data

	Training Accuracy (standard deviation)	Test Accuracy (standard deviation)
BAKER	96.0% (1.2%)	95.2% (0.9%)
SVM	96.0% (0.9%)	95.6% (1.7%)
GLM	96.0% (1.2%)	94.7% (1.3%)

The posterior predictive distribution for each sample in the test set is displayed in Figure 4.3. BAKER incorrectly classifies 8 samples 7 of which are benign. However, the 95%

credible intervals for each sample indicate that the uncertainty of some of the patients is much larger than that for some others. The credible intervals for a few incorrectly classified samples do not cover 0.5. It would be interesting to look more carefully at these samples, especially sample 8 which has very low posterior probability that strongly contradicts its class label .

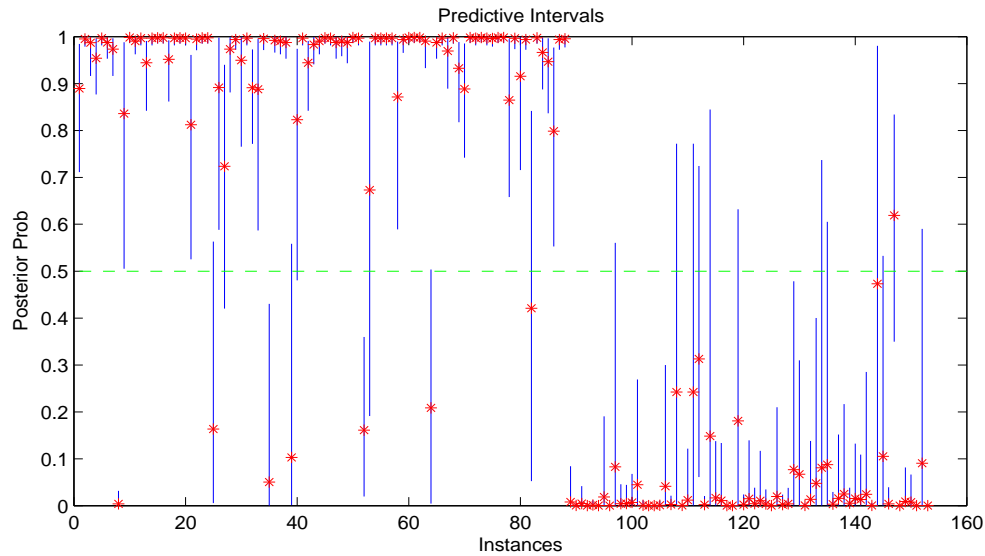


Figure 4.3: Posterior predictive probabilities for belonging to the benign class for 153 test samples. The blue line segments represent 95% credible intervals and the red star is the posterior mean. The first 88 samples are benign and the remaining 65 are malignant.

All 9 attributes seem relevant as reflected by the posterior means in Figure 4.4. The second variable “Uniformity of Cell Size” and the ninth variable “Mitoses” seem to be weaker than the others.

Johns Hopkins University Ionosphere database

The ionosphere data were examined in Sigillito et al. (1989) using a variety of neural networks. The class labels were good and bad. The input data had 34 attributes. They used the first 200 samples which included 100 bad and good samples as the training set and

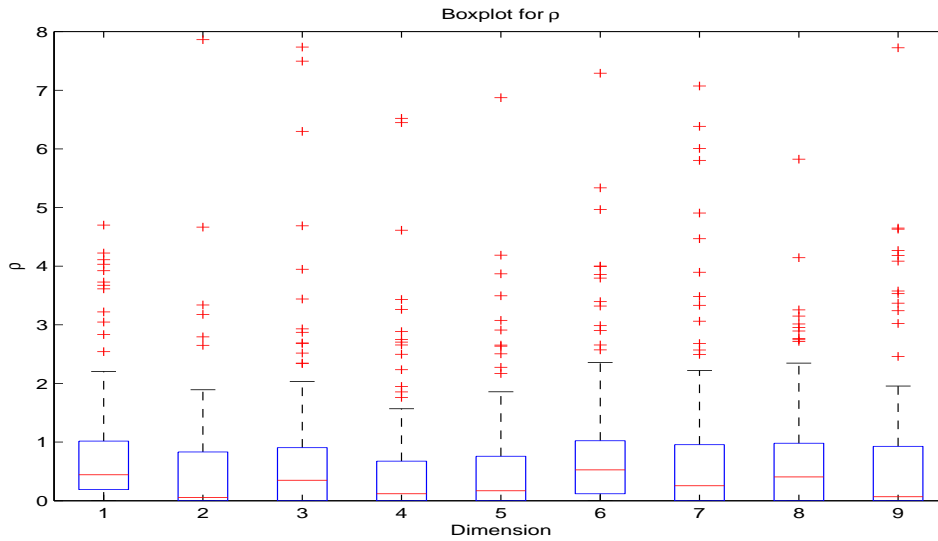


Figure 4.4: Boxplots of the posterior distribution of ρ_1, \dots, ρ_9 .

tested on the remaining 151 samples. They achieved test accuracies of 90.7% by a linear perceptron, 92% by a nonlinear perceptron and 96% by their implementation of backprop. We repeated this comparison using BAKER and SVM with a Gaussian kernel. The classifiers perform comparably as reflected in Table 4.3.

Table 4.3: The training and test accuracies for different methods on the Ionosphere data

	Training Accuracy (standard deviation)	Test Accuracy (standard deviation)
BAKER	97.5% (0.9%)	95.0% (1.9%)
SVM	98.9% (0.4%)	96.0% (1.1%)

The posterior predictive distribution on the test set as well as the posteriors for the relevance of each variable are plotted in Figure 4.5. Classifier is accurate on the test set but the uncertainty for some of the test samples is considerable. Variables 5, 9, 14, 23 have larger relative values and hence should be more significant.

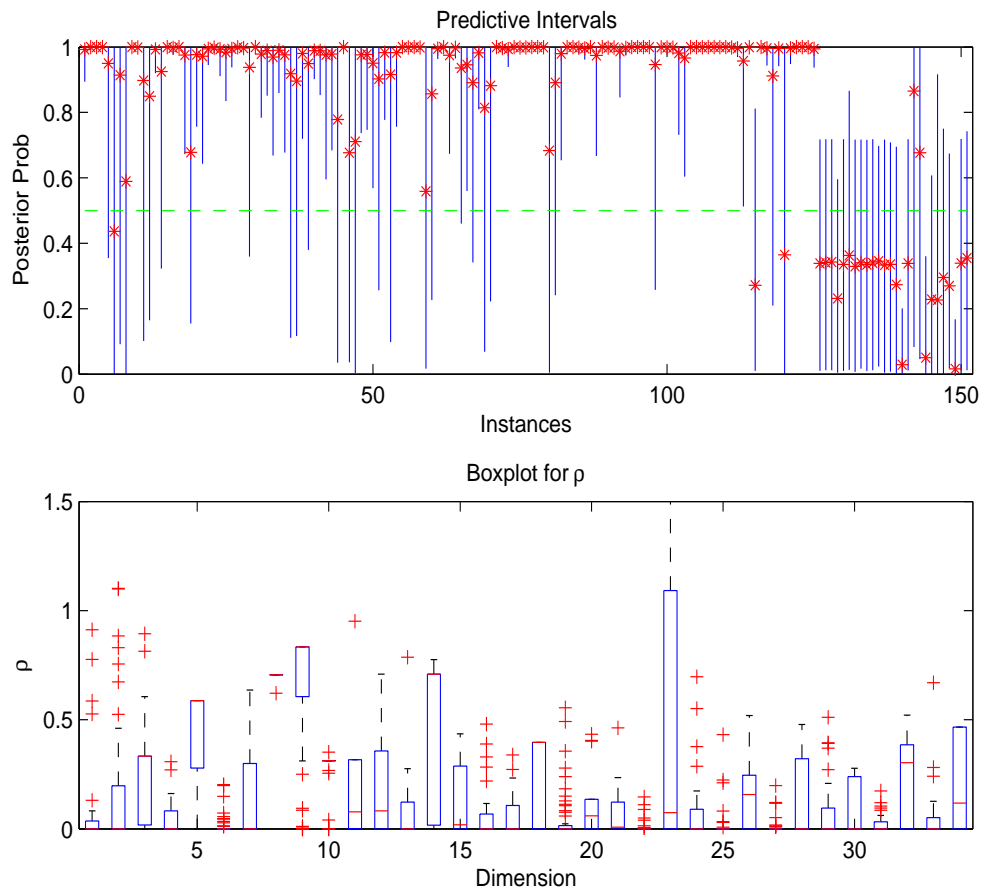


Figure 4.5: (a) The posterior predictive probability of a sample being good on the test samples. The first 125 samples are good and the remaining 75 are bad. Red stars are posterior means and blue lines are 95% credible intervals. (b) A boxplot showing the relevant significance for the explanatory variables.

4.5.4 Modeling considerations

The two most important considerations in Bayesian modeling are the computational efficiency of sampling from the posterior and the sensitivity of the posterior to prior specifications.

We first consider efficiency. The issue is the amount of time required to obtain coverage of the posterior distribution. The majority of time in the MCMC algorithm is in recomputing the kernel matrix for a new draw of ρ and the computation of the kernel decomposition. For example, a typical run on the Wisconsin Breast Cancer dataset with 1,000 iterations takes 153.00 seconds in which 91.2% or 139.54 seconds are spent on recomputing and decomposing the kernel (CPU: Intel Xeon 2×2.8GHz; Memory: 4096M; Matlab). This is the computational cost for feature selection.

The sensitivity of the posterior to prior specifications is another important issue. We have observed for the model without variable selection the posterior is relatively insensitive to the hyper-parameters. For the model with variable selection hyper-parameters have an effect on the posterior distribution of relevance for each variable, ρ . However, the prediction is robust with respect to the hyper-parameters.

In the model with variable selection the two considerations are the probability that a ρ is zero or not and the magnitude of a nonzero ρ . The parameter γ models the prior probability of a variable being relevant reflected by a Beta distribution. We use $a_\gamma = b_\gamma = 3$ to reflect a prior assumption of half the variables being relevant and the magnitude reflects some degree of concentration. Other values are also possible, but generally do not significantly influence the prediction. The posterior on the ρ values is more sensitive to prior specifications. This is due to the Gamma distribution draws from which can be very large. One possibility to address this is to use a truncated Gamma distribution. Another approach is to examine the hyper-parameters that control the tail of the Gamma distribution. The prior magnitude of ρ

depends on s which in turn depends on a_s . We examine the effect of a_s on the posterior by running the model setting $a_s = 0.5, 1, 5$. The comparisons are run on the Wisconsin Breast Cancer data set. The predictions are robust for all three settings, for example, we examine the first explanatory variable in the first instance of the test set. Figure (4.6) provides a boxplot of ρ_1 for the three settings of a_s . It is apparent that the posterior means of ρ_1 is very stable however the variance of the posterior distribution of ρ_1 is not. As a_s increases very large values of ρ_1 appear in the posterior distribution. These outliers drive the uncertainty in ρ . It is very important to consider this hyper-parameter when we model the uncertainty of variable relevance.

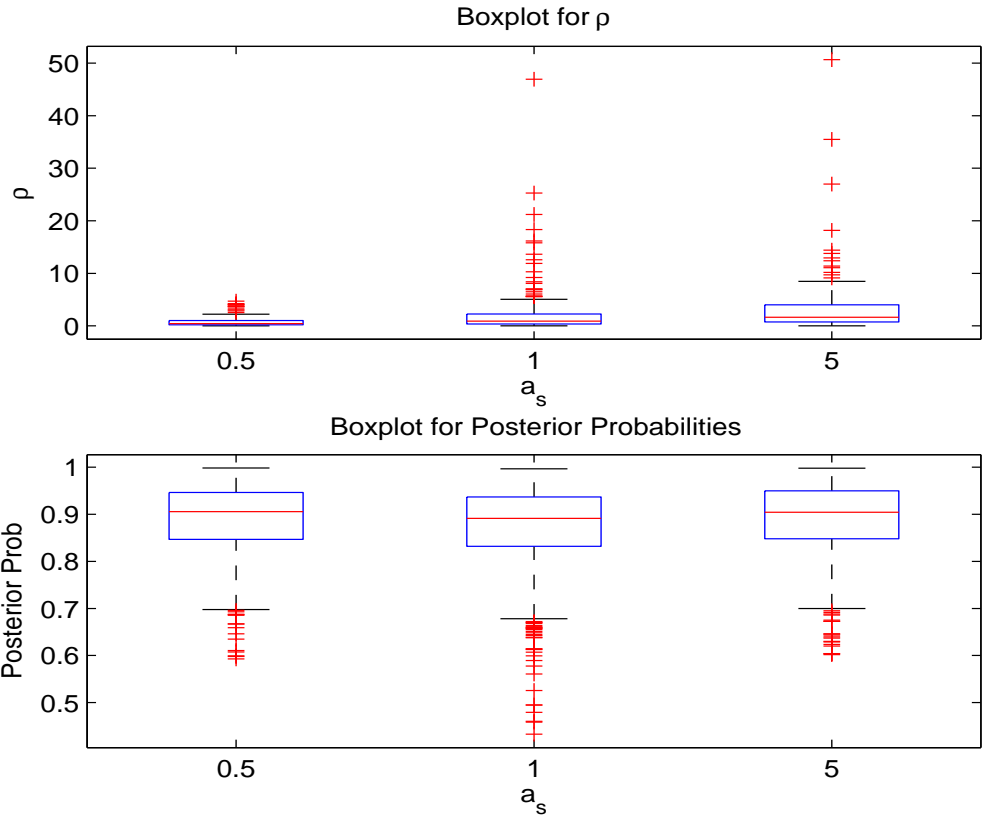


Figure 4.6: Boxplots for the posterior draws for ρ_1 (upper panel) and the posterior probability of the first benign sample under different hyper-parameter values of a_s . The predictions shown are similar, but the distribution of ρ_1 differs greatly due to outliers under larger a_s values.

4.6 Summary Comments

With the growth of interest in statistical classification and prediction methods in the machine learning communities, and an escalation of interest in applications among practitioners, there is a consequent need for refined theoretical understanding of the underlying statistical models as well as improved methodology and algorithms. We address each of these issues here. The theoretical foundation of our Bayesian kernel models is based on the equivalence between a class of functions induced by a nonparametric prior specification and a reproducing kernel Hilbert space. This Bayesian framework of the model allows for coherent inference, assessment of uncertainty, and access to the posterior distributions via Markov chain Monte Carlo sampling. Practical issues such as choice of hyper-parameters and variable selection are automatically incorporated into the Bayesian modeling and inference.

The Bayesian kernel model suggests several interesting future directions as well as open problems. The computational challenges of searching high-dimensional parameter space is of utmost importance and for variable selection increasing the efficiency of the MCMC to be able to handle thousands of variables is an open problem of great practical importance. The nonparametric Bayesian kernel model we developed is an example of a more general framework described in Pillai et al. (2007). Further exploration of other process priors from a theoretical, computational, and applied data analysis perspective is of interest.

A striking example of the flexibility and coherence of the Bayesian kernel model is its application to what is referred to as the semi-supervised problem in the machine learning literature, the incorporation of unlabeled data – an example of ancillary design data – in classification and regression problems. Our Bayesian kernel model incorporates the unlabeled data in a natural way without having to introduce additional penalties to the loss function as is the case for regularization approaches, which is discussed in detail in Liang

et al. (2007).

Appendix

In this appendix, we introduce the properties of the Grassmann manifold that allow for an intuition of the meaning of the distance metric in section 2.2.3. We first develop the concepts of the Riemannian metric and geodesic distance. We then develop the definition of the geodesic distance on the Grassmann manifold. For more information on the differential geometry of the Grassmann manifold see Kobayashi and Nomizu (1996).

Given a differentiable manifold \mathcal{M} a Riemannian metric is an inner product function

$$g_p(\xi, \eta) \equiv \langle \xi, \eta \rangle_p : T_p\mathcal{M} \times T_p\mathcal{M} \rightarrow \mathbb{R},$$

where ξ, η are tangent vectors of the manifold \mathcal{M} at a point $p \in \mathcal{M}$ and for $\xi, \eta, \nu \in T_p\mathcal{M}$ and $a, b \in \mathbb{R}$

$$\begin{aligned} g_p(a\xi + b\eta, \nu) &= ag_p(\xi, \nu) + bg_p(\eta, \nu) \\ g_p(\nu, a\xi + b\eta) &= ag_p(\nu, \xi) + bg_p(\nu, \eta) \\ g_p(\nu, \xi) &= g_p(\nu, \xi) \end{aligned}$$

and for every $\xi \in T_p\mathcal{M}$ there exists $\nu \in T_p\mathcal{M}$ such that $g_p(\xi, \nu) \neq 0$. Also given an open subset U of the manifold and tangent vectors ν and ξ on U the function $g_p(\xi, \nu)$ is a smooth function of p .

A geodesic curve on a smooth manifold is the shortest curve between two points as well as the straightest curve between two points. It is useful to think of curves on manifolds as a map $c : [0, 1] \rightarrow \mathcal{M}$ where the interval $[0, 1]$ is thought of as time so the curve is a function of time, $c(t)$. For any point p on the manifold and any tangent vector $\nu \in T_p\mathcal{M}$ there exists

a unique geodesic curve γ such that

$$(1) \gamma(0) = p, \quad (2) \dot{\gamma}(0) = \nu, \quad (3) \nabla_{\dot{\gamma}} \dot{\gamma} = 0, \quad (.8)$$

where $\dot{\gamma}$ is the derivative with respect to t and in this case is a tangent vector on the manifold, $\dot{\gamma} \in T_p \mathcal{M}$. Statement (1) is that the curve starts at point p . Statement (2) is the curve initially moves in direction ν with velocity $|\nu|$. In statement (3) ∇ is the affine connection on the manifold which is a generalization of the standard notion of a directional derivative. For our purposes, interpret $\nabla_V f$ as the differentiation of a f in the direction V . In the case of statement (3) this means that at points on the geodesic curve the velocity of the geodesic curve in the direction $\dot{\gamma}$ is constant, there is no acceleration. The definition of a geodesic curve in (.8) is equivalent to a second order ordinary differential equation, so geodesics are uniquely defined by an initial point and an initial direction. This is why the shortest curve is the straightest.

For a Riemannian manifold \mathcal{M} with Riemannian metric g the length of a curve $c : [0, 1] \rightarrow \mathcal{M}$ that is differentiable with velocity vector \dot{c} is

$$L_0^1(c) \equiv \int_0^1 \sqrt{g(\dot{c}(t), \dot{c}(t))} dt.$$

The distance function $d : \mathcal{M} \times \mathcal{M} \mathbb{R}^+$ is defined as

$$d(p, q) = \inf L(c)$$

where the infimum is over all differentiable curves such that $c(0) = p$ and $c(1) = q$. The curve that minimizes this distance is the geodesic γ with $\gamma(0) = p$ and $\gamma(1) = q$. This relates the geodesic to the Riemannian metric.

We now develop the intuition behind the distance function on the Grassmann manifold.

Let $\mathcal{G}_{(d,p)}$ be the Grassmann manifold, a point \mathcal{U} on the manifold is a d dimensional subspace of \mathbb{R}^p . A matrix $U \in \mathbb{R}^{p \times d}$ can be specified by basis vectors of the subspace such that $\text{span}(U) = \mathcal{U}$ and $U = (u_1, \dots, u_d)$. The set of matrices that span \mathcal{U} is the noncompact Stiefel manifold

$$\mathcal{S}_{(d,p)} \equiv \{U \in \mathbb{R}^{p \times d} : \text{rank}(U) = d\}.$$

For each subspace \mathcal{Z} there exists an equivalence class of matrices \mathcal{Z} such that $\text{span}(Z) = \mathcal{Z}$ for each $Z \in \mathcal{Z}$. However, one can single out a single unique matrix as follows. Let $W \in \mathcal{S}_{(d,p)}$ we define the affine cross section as

$$\mathcal{C}_W \equiv \{Z \in \mathcal{S}_{(d,p)} : W'(Z - W) = 0\}.$$

If $W'Z$ is invertible then \mathcal{Z} intersects \mathcal{C}_W at the point $Z(W'Z)^{-1}W'W$, otherwise they do not intersect. We define the set

$$\mathcal{Z}_W \equiv \{\text{span}(Z) : W'Z \text{ is invertible}\}.$$

We define the maps

$$\sigma_W : \mathcal{Z}_W \rightarrow \mathcal{C}_W, \quad \pi(W) : W \rightarrow \text{span}(W).$$

Define as \mathcal{W} a subspace and W as its basis, the tangent vector ξ at a point \mathcal{W} on the Grassmann manifold is $T_{\mathcal{W}}\mathcal{G}_{(d,p)}$. The matrix W can be adapted to capture the directions of W that have an effect on changing the subspace \mathcal{W} . Define

$$\xi_W = d\sigma_W(\mathcal{W})\xi, \quad d\pi(W)\xi_W = \xi,$$

where ξ_W projects to the tangent ξ via π . The representation ξ_W is called the horizontal lift

of $\xi \in T_{\mathcal{W}}\mathcal{G}_{(d,p)}$ and are the directions of W which if varied will change the subspace \mathcal{W} .

In the case of the Grassmann manifold the Riemannian metric is

$$\langle \xi, \eta \rangle_{\mathcal{Z}} = \text{Tr} \left((Z'Z)^{-1} \xi'_Z \eta_Z \right),$$

where $\text{span}(Z) = \mathcal{Z}$ and the metric does not depend on the basis that spans \mathcal{Z} . Given the above metric the geodesic $\gamma(t)$ between two points \mathcal{J} and \mathcal{Z} can be defined by the initial tangent vector $\xi \in T_{\mathcal{J}}\mathcal{G}_{(d,p)}$ where $\gamma(0) = \mathcal{J}$, $\gamma(1) = \mathcal{Z}$, and $\dot{\gamma}(0) = \xi$. Given J and Z that span \mathcal{J} and \mathcal{Z} respectively, the shortest curve from \mathcal{J} to \mathcal{Z} on $\mathcal{G}_{(d,p)}$ is

$$\begin{aligned} \xi_J &= U\Theta V \\ U\Sigma V' &= (I - J(J'J)^{-1}J')Z(J'Z)^{-1} \\ \Theta &= \text{atan}(\Sigma). \end{aligned}$$

and

$$\text{dist}(\mathcal{J}, \mathcal{Z}) = \sqrt{\text{Tr}(\Theta^2)}$$

Bibliography

- Absil, P.-A., R. Mahony, and R. Sepulchre (2004). Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematicae* 80, 199–220.
- Belkin, M. and P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15(6), 1373–1396.
- Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press.
- Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* 1, 353–355.
- Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Wadsworth.
- Chapelle, O., V. Vapnik, O. Bousquet, and S. Mukherjee (2002). Choosing multiple parameters for support vector machines. *Machine Learning* 46(1-3), 131–159.
- Chen, M., J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin (2009). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. Technical report.
- Cook, R. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* 91, 983–992.
- Cook, R. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science* 22(1), 1–26.
- Cook, R. and S. Weisberg (1991). Discussion of "sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.* 86, 328–332.

- Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning* 20(3), 273–297.
- Cox, T. F. and M. Cox (2001). *Multidimensional Scaling*. Chapman and Hall.
- Diaconis, P. and D. Freedman (1980). De Finetti’s theorem for markov chains. *The Annals of Probability* 8(1), 115–130.
- Donoho, D. and C. Grimes (2003). Hessian eigenmaps: new locally linear embedding techniques for highdimensional data. *Proceedings of the National Academy of Sciences* 100, 5591–5596.
- Donoho, D. and I. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425–455.
- Dunson, D. B. and J. Park (2008). Kernel stick-breaking processes. *Biometrika* 89, 268–277.
- Dunson, D. B., X. Ya, and C. Lawrence (2008). The matrix stick-breaking process: Flexible Bayes meta-analysis. *J. Amer. Statist. Assoc.* 103, 317–327.
- Edelman, E., J. Guinney, J. Chi, P. Febbo, and S. Mukherjee. Modeling cancer progression via pathway dependencies. *PLoS Comp. Bio.*
- Edelman, E., A. Porello, J. Guinney, B. Balakumaran, A. Bild, P. Febbo, and S. Mukherjee (2006). Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 22(14), 108–116.
- Escobar, M. and M. West (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* 90, 577–588.

- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1(2), 209–230.
- Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* 2(4), 615–629.
- Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics* 7(II), 179–188.
- Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* 19(1), 1–141.
- Friedman, J. H. and W. Stuetzle (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* 76, 817–823.
- Gelfand, A., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* 100(471), 1021–1035.
- Globerson, A. and S. Roweis (2006). Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems 18*, pp. 451–458.
- Goldberger, J., S. Roweis, G. Hinton, and R. Salakhutdinov (2005). Neighbourhood component analysis. In *Advances in Neural Information Processing Systems 17*, pp. 513–520.
- Griffin, J. and M. Steel (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* 101, 179–194.
- Hastie, T. and R. Tibshirani (1996a). Discriminant adaptive nearest neighbor classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 607–615.
- Hastie, T. and R. Tibshirani (1996b). Discriminant analysis by Gaussian mixtures. *J. Roy. Statist. Soc. Ser. B* 58(1), 155–176.

- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer.
- He, X. and P. Niyogi (2003). Locality preserving projections. In *Advances in Neural Information Processing Systems 16*. MIT Press.
- Iorio, M. D., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An anova model for dependent random measures. *J. Amer. Statist. Assoc.* 99, 205–215.
- Ishwaran, H. and L. James (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* 96(453), 161–173.
- Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* 30(5), 509–541.
- Kendall, W. S. (1990). Probability, convexity and harmonic maps with small image. i. uniqueness and fine existence. *Proc. London Math. Soc.* 61(2), 371–406.
- Kimeldorf, G. and G. Wahba (1971). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* 41(2), 495–502.
- Kobayashi, S. and K. Nomizu (1996). *Foundations of Differential Geometry*, Volume 1. Wiley-Interscience.
- Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.
- Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modeling. *Ann. Statist.* 20, 1222–1235.
- Lavine, M. (1994). More aspects of Pólya tree distributions for statistical modeling. *Ann. Statist.* 22, 1161–1176.

- Lenk, P. (1988). The logistic normal distribution for Bayesian nonparametric predictive densities. *J. Amer. Statist. Assoc.* 83, 509–516.
- Li, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* 86, 316–342.
- Li, K. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. *Ann. Statist.* 97, 1025–1039.
- Liang, F., S. Mukherjee, and M. West (2007). The use of unlabeled data in predictive modeling. *Statistical Science* 22(2), 189–205.
- Liao, M. (2005). *Bayesian Models and Machine Learning with Gene Expression Analysis Applications*. Ph. D. thesis.
- Lopes, H. F. and M. West (2004). Bayesian model assessment in factor analysis. *statistica sinica* 14, 41–67.
- MacEachern, S. and P. Müller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* 7, 223–238.
- MacEachern, S. N. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55.
- Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. Lond. Ser. A* 209, 415–446.
- Mika, S., B. Schölkopf, A. Smola, K. Müller, M. Scholz, and G. Rätsch (1999). Kernel PCA and de-noising in feature spaces. In *Advances in Neural Information Processing Systems 11*.

- Mukherjee, S., P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. Golub, and J. Mesirov (2003). Estimating dataset size requirements for classifying DNA Microarray data. *Journal of Computational Biology* 10, 119–143.
- Mukherjee, S. and Q. Wu (2006). Estimation of gradients and coordinate covariation in classification. *J. Mach. Learn. Res.* 7, 2481–2514.
- Mukherjee, S. and D. Zhou (2006). Learning coordinate covariances via gradients. *J. Mach. Learn. Res.* 7, 519–549.
- Mukherjee, S., D.-X. Zhou, and Q. Wu (2006). Learning gradients and feature selection on manifolds. Technical report, ISDS, Duke University.
- Müller, P., F. Quintana, , and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *J. Amer. Statist. Assoc.* 66, 735–749.
- Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics* 27, 105–126.
- Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics* 26, 373–393.
- Petrone, S. and L. Wasserman (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc. Ser. B* 64, 79–100.
- Pillai, N., Q. Wu, F. Liang, S. Mukherjee, and R. Wolpert (2007). Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res.* 8, 1769–1797.
- Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences U.S.A.* 98, 149–54.

- Roweis, S. and L. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326.
- Schölkopf, B. and A. J. Smola (2001). *Learning with Kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge: The MIT Press.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *statistica sinica* 4, 639–650.
- Sigillito, V., S. Wing, L. Hutton, and K. Baker (1989). Classification of radar returns from the ionosphere using neural networks. Technical Report 10, Johns Hopkins APL Technical Digest.
- Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning* 46(1-3), 21–52.
- Speed, T. and H. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* 14, 138–150.
- Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *J. Mach. Learn. Res.* 8, 1027–1061.
- Tenenbaum, J., V. de Silva, and J. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323.
- Tikhonov, A. N. and V. Y. Arsenin (1977). *Solution of Ill-posed Problems*. Washington: Winston Sons.
- Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* 1, 211–244.

- Tokdar, S., Y. Zhu, and J. Ghosh (2008). A Bayesian implementation of sufficient dimension reduction in regression. Technical report, Carnegie Mellon University.
- Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.
- Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.
- West, M. (2003). Bayesian factor regression models in the “large p , small n ” paradigm. In *Bayesian Statistics*, pp. 723–732. Oxford University Press.
- Wu, Q., J. Guinney, M. Maggioni, and S. Mukherjee (2007). Learning gradients: Predictive models that infer geometry and dependence. Technical report, ISDS Discussion Paper, Duke University.
- Wu, Q., F. Liang, and S. Mukherjee (2007). Regularized sliced inverse regression for kernel models. Technical report, ISDS, Duke University.
- Wu, Q., S. Mukherjee, and F. Liang (2008). Localized sliced inverse regression. In *NIPS*, pp. 1785–1792.
- Xia, Y., H. Tong, W. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* 64(3), 363–410.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc.* 81, 446–451.
- Zhang, J. (1992). Selecting typical instances in instance-based learning. In *Proceedings of the Ninth International Machine Learning Conference*, pp. 470–479. Morgan Kaufmann Publishers, San Francisco.

Biography

Kai Mao obtained the Bachelor of Science degree in Mathematics at Tsinghua University in 2005. He has since been at Duke University in the Ph.D. program of Statistical Science. He obtained the Master of Science degree in Statistics at Duke University in 2008.