Nonparametric Bayes regression and classification through mixtures of product kernels

David B. Dunson & Abhishek Bhattacharya

Department of Statistical Science

Box 90251, Duke University

Durham, NC 27708-0251, USA

dunson@stat.duke.edu

Summary

It is routine in many fields to collect data having a variety of measurement scales and supports. For example, in biomedical studies for each patient one may collect functional data on a biomarker over time, gene expression values normalized to lie on a hypersphere to remove artifacts, clinical and demographic covariates and a health outcome. A common interest focuses on building predictive models, with parametric assumptions seldom supported by prior knowledge. Hence, it is most appropriate to define a prior with large support allowing the conditional distribution of the response given predictors to be unknown and changing flexibly across the predictor space not just in the mean but also in the variance and shape. Building on earlier work on Dirichlet process mixtures, we describe a simple and general strategy for inducing models for conditional distributions through discrete mixtures of product kernel models for joint distributions of predictors and response variables. Computation is straightforward and the approach can easily accommodate combining of widely disparate data types, including vector data in a Euclidean space, categorical observations, functions, images and manifold data.

Key Words: Clustering; Data fusion; Density regression; Joint modeling; Latent class; Missing data; Object data; Transfer learning.

1 Introduction

Consider the general problem of predicting a response $Y \in \mathcal{Y}$ based on predictors $X \in \mathcal{X}$, where $\mathcal{Y}$ and $\mathcal{X}$ are initially considered to be arbitrary metric spaces. From an applied perspective, we are motivated by the need to accommodate data having a rich variety of measurement scales and supports, as it is increasingly common to collect multivariate and disparate data in modern biomedical studies as well as in other areas. For example, for each study subject we may obtain information on a categorical response variable $Y \in \{1, \ldots, c\}$ as well as predictors having different supports including categorical, Euclidean, a hypersphere and a planar shape space. In other applications, the response may be multivariate and may have constrained support and the predictors may include functional data and images. It is not straightforward to combine such disparate and multidimensional data in building flexible models for classification and regression, while defining a general framework that can be easily adapted to allow a rich variety of data structures and incorporate additional data as they become available (e.g., from an additional assay run on samples for one or more subjects under study). The focus of this article is on defining a very general nonparametric Bayes modeling framework for the conditional distribution of Y given $X = x$ through joint modeling of $Z = (Y, X)$.

The idea of inducing a flexible model on the conditional of Y given $X = x$ through a flexible model for the joint is not new. In the setting in which $\mathcal{Y} = \Re$ and $\mathcal{X} = \Re^p$, [19] proposed to induce a flexible model for $\mathrm{E}(Y \mid X = x)$ through a Dirichlet process (DP) [13, 14] mixture of multivariate Gaussian distributions for $Z = (Y, X')'$. Such a model induces a prior with *large support* on the conditional density of Y given $X = x$. Large support in this context means that the prior can generate conditional densities, $\{f(y|x), y \in \Re, x \in \Re^p\}$, that are arbitrarily close to any true data-generating conditional density, $\{f_0(y|x), y \in \Re, x \in \Re^p\}$, in a large class. From a practical perspective, the ramifications are that we

can characterize flexible relationships between X and Y not only in the mean $\mathrm{E}(Y \mid X = x)$ but also in other aspects of the conditional density including the variance, shape and quantiles. A flexible model for the conditional density having large support will automatically allow the quantiles of $f(y|x)$ to have differing nonlinear relationships with the predictors. In contrast, most approaches for nonparametric regression model $\mathrm{E}(Y \mid X = x)$ flexibly while making restrictive assumptions about the residual density, such as homoscedasticity. Such assumptions typically do not arise out of prior knowledge and can lead to misleading inferences and predictions, particularly when the focus is not just on the mean response but also on the tails.

Before considering generalizations of the [19] approach to non-Euclidean spaces, it is useful to discuss some of the specifics of the model in the simple case. Letting $z_i = (y_i, x_i')'$ denote the data for subjects $i = 1, \ldots, n$, the DPM of multivariate Gaussians model for the density of z_i implies that

$$
\begin{aligned}
(y_i \mid x_i, S_i) &\sim \mathrm{N}(x_i'\beta_{S_i}, \tau_{S_i}), \\
(x_i \mid S_i) &\sim \mathrm{N}_p(\mu_{S_i}, \Sigma_{S_i}),
\end{aligned}
\tag{1}
$$

where S_i is a cluster index for subject i and, for subjects in cluster h, β_h are regression coefficients, τ_h is the response model residual variance, μ_h is the predictor mean, and Σ_h is the predictor covariance. The prior on $S = (S_1, \ldots, S_n)$ follows the [5] Pólya urn scheme, with the DP precision parameter α controlling the tendency to allocate subjects to many clusters; for small α the prior favors allocation to few clusters.

It follows from (1) and [5] that the predictive density of Y for a new subject having predictor values x_{n+1} is

$$
f(y_{n+1} \mid x_{n+1}) \approx \sum_{h=1}^{k} \pi_h(x_{n+1}) N(y; x_{n+1}'\beta_h, \tau_h),
\tag{2}
$$

where $k = \max\{S_1, \ldots, S_n\}$ is the number of cluster in the n subjects, the approximation

3

assumes that $\alpha/(\alpha + n) \approx 0$, and the predictor-dependent weight on component h is

$$\pi_h(x_{n+1}) = \frac{n_h N_p(x_{n+1}; \mu_h, \Sigma_h)}{\sum_{l=1}^{k} n_l N(x_{n+1}; \mu_l, \Sigma_l)}, \quad h = 1, \ldots, k, \tag{3}$$

where $n_h = \sum_{i=1}^{n} 1(S_i = h)$ is the number of subjects in cluster h. Hence, from (2) - (3), the conditional density of y_{n+1} given x_{n+1} is modeled as a predictor-dependent mixture of normal linear regression models. This is related to the hierarchical mixture-of-experts model [17], but instead of conditioning on the predictors and fitting a flexible model to $\pi_h(x)$ the weights arise as a natural consequence of the joint DPM model for y and x.

This approach of inducing a flexible kernel-weighted mixture of normal linear regressions for $f(y|x)$ through a *joint* DPM of Gaussians for y and x has considerable conceptual appeal. However, difficulties arise in implementation for moderate to large p, as it is then necessary to estimate a $p \times p$ covariance matrix specific to each component. [23] recently proposed a modification for classification problems in which $y \in \{1, \ldots, c\}$ and $x \in \Re^p$. They replaced the normal linear regression model in the first line of (1) with a multinomial logit model, while assuming Σ_h in line 2 is diagonal. [20] instead considered a general class of joint DPM models with $f(z_i|\theta_i) = f_1(y_i \mid x_i, \varphi_i) f_2(x_i \mid \gamma_i)$, where $\theta_i = (\varphi_i, \gamma_i) \sim P$ and $P \sim DP(\alpha P_0)$. From this, they derived a predictor-dependent generalization of the [5] DP prediction rule.

In considering general applications, an additional level of computational complexity is often added in allowing dependence between the different elements of x_i and between y_i and x_i *within* each component. Hence, there is practical appeal in consider DP mixtures of independent (or product) kernels. For example, in the special case in which $y_i \in \Re$ and $x_i \in \Re^p$, we could replace line 1 of (1) with $(y_i \mid S_i) \sim N(\psi_{S_i}, \tau_{S_i})$ while assuming Σ_h is diagonal for all h in line 2. In this article, we will propose a very general class of discrete mixtures of product kernels and will provide a detailed discussion of the advantages and disadvantages of models in this class.

Section 2 describes the proposed class of models and discusses properties from a practical

perspective, while citing theoretical results in a companion paper [1]. Section 3 provides some illustrative examples. Section 4 discusses some drawbacks of the general strategy of fitting joint models when the interest is in the conditional, while describing some important future directions.

2 Discrete Mixtures of Product Kernels

2.1 Model Description

Suppose that $Y \in \mathcal{Y}$ and $X = \{X_1, \ldots, X_p\}$ with $X_j \in \mathcal{X}_j$, for $j = 1, \ldots, p$. We let the sample spaces $\mathcal{Y}, \mathcal{X}_1, \ldots, \mathcal{X}_p$ be very general ranging from subsets of $\Re$ or $\{1, 2, \ldots, \infty\}$ to arbitrary non-Euclidean manifolds, such as the hypersphere. Letting y_i and $x_i = \{x_{i1}, \ldots, x_{ip}\}$ denote the response and predictor values for subject i and assuming $(y_i, x_i) \overset{iid}{\sim} f$, for $i = 1, \ldots, n$, we let

$$f(y, x) = \int \left\{ \mathcal{K}^{(y)}(y; \theta^{(y)}) \prod_{j=1}^{p} \mathcal{K}^{(x_j)}(x_j; \theta^{(x_j)}) \right\} dP(\theta), \quad \theta = \{\theta^{(y)}, \theta^{(x_1)}, \ldots, \theta^{(x_p)}\}, \quad (4)$$

where $K^{(y)}$ is a parametric density on $\mathcal{Y}$, $\mathcal{K}^{(x_j)}$ is a parametric density on $\mathcal{X}_j$, for $j = 1, \ldots, p$, and P is a mixing measure assigned a prior $\mathcal{P}$. In particular, we assume $\mathcal{P}$ is chosen so that

$$P = \sum_{h=1}^{k} \pi_h \delta_{\Theta_h}, \quad \Theta_h = \{\Theta_h^{(y)}, \Theta_h^{(x_1)}, \ldots, \Theta_h^{(x_p)}\} \sim P_0 = P_0^{(y)} \prod_{j=1}^{p} P_{0j}^{(x)}, \quad (5)$$

where P_0 is a base measure, which is constructed as a product, and k can be either finite or infinite. Prior (5) encompasses a broad class of species sampling priors, with the Dirichlet process and two parameter Poisson-Dirichlet (Pitman-Yor) process arising as special cases. The Dirichlet process is obtained by letting $k = \infty$ and $\pi_h = V_h \prod_{l<h}(1 - V_l)$ with $V_h \sim$ beta$(1, \alpha)$ independently for $h = 1, \ldots, \infty$.

Model (4) - (5) implies the following model for the conditional density $f(y|x)$,

$$f(y|x, \pi, \Theta) = \sum_{h=1}^{k} \left\{ \frac{\pi_h \prod_{j=1}^{p} \mathcal{K}^{(x_j)}(x_j; \Theta_h^{(x_j)})}{\sum_{l=1}^{k} \pi_l \prod_{j=1}^{p} \mathcal{K}^{(x_j)}(x_j; \Theta_l^{(x_j)})} \right\} \mathcal{K}^{(y)}(y; \Theta_h^{(y)})$$

$$= \sum_{h=1}^{k} \pi_h(x) \mathcal{K}^{(y)}(y; \Theta_h^{(y)}), \tag{6}$$

which expresses the conditional density as a predictor-dependent mixture of kernels that do not depend on x. As illustration, consider the simple example in which $p = 1$, $\mathcal{X}_1 = \Re$, $\mathcal{Y} = \Re$, and we choose Gaussian kernels. Then, we have

$$f(y|x, \pi, \Theta) = \sum_{h=1}^{k} \left\{ \frac{\pi_h N(x; \mu_h, \sigma_h^2)}{\sum_{l=1}^{k} \pi_l N(x; \mu_l, \sigma_l^2)} \right\} N(y; \psi_h, \tau_h), \tag{7}$$

One can think of $N(\psi_h, \tau_h)$, for $h = 1, \ldots, k$, as basis densities, with the conditional densities expressed as convex combinations of these bases. Hence, the conditional densities $f(y|x)$ at different x values are expressed as a mixture of a common collection of normal basis distributions. The probability weights vary smoothly with x, with the weights $\pi(x) = \{\pi_1(x), \ldots, \pi_k(x)\}$ and $\pi(x') = \{\pi_1(x'), \ldots, \pi_k(x')\}$ converging as $x \to x'$.

It is interesting that such a rich model can be induced through the very simple structure in (4) - (5), which does not directly model dependence between Y and X or between the different elements of X. In fact, it can be shown that the dependence only comes in through sharing of a common cluster allocation latent class variable across the different data types. Such shared latent class models are useful not only in modeling of conditional distributions in regression and classification but also in data fusion and combining of information from disparate data sources.

For data $\{y_i, x_i\}$ generated independently from the joint density $f(y, x)$ described in (4) - (5), we have

$$
\begin{aligned}
y_i &\sim \mathcal{K}^{(y)}(\Theta_{S_i}^{(y)}), \quad \Theta_h^{(y)} \sim P_0^{(y)}, \\
x_{ij} &\sim \mathcal{K}^{(x_j)}(\Theta_{S_i}^{(x_j)}), \quad \Theta_h^{(x_j)} \sim P_{0j}^{(x)}, \ j = 1, \ldots, p \\
S_i &\sim \sum_{h=1}^{k} \pi_h \delta_h,
\end{aligned}
\tag{8}
$$

where δ_h denotes a degenerate distribution with all its mass at the integer h. Hence, to sample from the proposed product kernel mixture model, we simply generate cluster indices (latent

classes) $S_1, \ldots, S_n$ independently from a multinomial-type distribution. Then, conditionally on the latent class status for the different subjects, the response and different predictors are independent with the parameters in the different likelihoods assigned independent priors. As will be made clear in the next section, this conditional independence greatly facilitates posterior computation in very general problems involving mixtures of different complicated and high-dimensional data types.

2.2 Posterior Computation

To illustrate posterior computation for discrete mixtures of product kernel models, we focus on the simple case in which

$$\pi = (\pi_1, \ldots, \pi_k)' \sim \text{Diri}(a_1, \ldots, a_k). \tag{9}$$

Generalizations to accommodate $k = \infty$ are straightforward using recently-developed algorithms described in [26, 18]. By letting $a_h = \alpha/k$, the finite Dirichlet prior can be used as an approximation to the Dirichlet process [16], which improves in accuracy as k increases. In this case, k is not the number of mixture components occupied by the n subjects in the sample, but is instead an upper bound on the number of components. For sufficiently large values of k, the choice of k does not make a practical difference in the analysis. [22] recently showed that when the data are generated from a finite mixture model with k_0 components, one can obtain posterior consistency in using a finite mixture model with $k > k_0$ under some weak conditions on the prior in (9). This is due to the tendency to effectively delete components through having posterior distributions for π_h that are increasingly concentrated near zero for unnecessary components.

We recommend the approach of monitoring the number of occupied components $k_n = \sum_{h=1}^{k} 1(n_h > 0)$ with $n_h = \sum_{i=1}^{n} 1(S_i = h)$, for $h = 1, \ldots, k$, across the MCMC iterations. If any of the samples of k_n after burn-in are within a couple units of k, then the upper

bound k should be increased. This can potentially be implemented with an adaptive MCMC algorithm designed to satisfy the diminishing adaptation condition [21]. In our experience, we find that mixing is often better in using (9) instead of a stick-breaking representation in which the mixture components are non-exchangeable and hence there is greater sensitivity to starting values.

A simple data augmentation MCMC algorithm can proceed through the following sampling steps:

1. Update the cluster allocation S_i for each subject by sampling from the conditional posterior with

$$\Pr(S_i = h \,|\, -) = \frac{\pi_h \mathcal{K}^{(y)}(y_i; \Theta_h^{(y)}) \prod_{j=1}^{p} \mathcal{K}^{(x_j)}(x_{ij}; \Theta_h^{(x_j)})}{\sum_{l=1}^{k} \pi_l \mathcal{K}^{(y)}(y_i; \Theta_l^{(y)}) \prod_{j=1}^{p} \mathcal{K}^{(x_j)}(x_{ij}; \Theta_l^{(x_j)})}, \quad h = 1, \ldots, k, \quad (10)$$

which is easy to calculate quickly. The probability of allocation to cluster h is proportional to the prior probability on cluster h multiplied by the conditional likelihood of the data $\{y_i, x_i\}$ given allocation. Hence, allocation to clusters is driven by improving the fit of not only the conditional likelihood of the response given the predictors but also the predictor likelihood. In certain cases this can present practical problems, as when many clusters are introduced to better fit the x likelihood but these clusters are not needed for characterizing $f(y|x)$. Such pitfalls of the joint modeling approach are discussed further in Section 4.

2. Update the weights on each component from the conjugate conditional posterior

$$(\pi \,|\, -) \sim \text{Diri}(a_1 + n_1, \ldots, a_k + n_k). \quad (11)$$

3. Update the response parameters $\Theta_h^{(y)}$ specific to each cluster $h = 1, \ldots, k$ from

$$(\Theta_h^{(y)} \,|\, -) \propto P_0^{(y)}(\Theta_h^{(y)}) \prod_{i:S_i=h} \mathcal{K}^{(y)}(y_i; \Theta_h^{(y)}). \quad (12)$$

8

Often, $P_0^{(y)}$ can be chosen to be conjugate so that this conditional is available in a simple form that can be sampled from directly. This is one practical advantage of using the product kernel mixture formulation under conditional independence. If the conditional is non-conjugate, Metropolis-Hasting can be used.

4. Similarly, update the predictor parameters $\Theta_h^{(x_j)}$ for $j = 1, \ldots, p$ and $h = 1, \ldots, k$ from

$$(\Theta_h^{(x_j)} \mid -) \propto P_{0j}^{(x)}(\Theta_h^{(x_j)}) \prod_{i:S_i=h} \mathcal{K}^{(x_j)}(x_{ij}; \Theta_h^{(x_j)}). \tag{13}$$

These simple steps should be repeated a large number of times, with a burn-in discarded to allow convergence. Due to the well-known label switching problem, one should not assess convergence and mixing or calculate posterior summaries of the mixture component-specific parameters (the Θ_h's) without applying post-processing or some related approach as described in [24]. Our own view is that it is typically misleading to attempt to interpret clusters and mixture component-specific parameters, since the posterior on these quantities is extremely sensitive to the choice of kernels and specific conditional independence assumptions made. Instead, one can use the mixture model and clustering simply as a tool for generating an extremely flexible model for the joint distribution of Y and X and for the conditional of Y given $X = x$.

An appealing aspect of the joint modeling approach is that it is trivial to accommodate missing data under a missing at random assumption. If subject i is missing some of the measurements (this can be a subset of the predictors and/or the response), then one simply modifies the conditional probability of $S_i = h$ in step 1 above to update π_h with the likelihood for only those data measured for subject i. Then, in (12) - (13) one modifies $\prod_{i:S_i=h}$ to remove subjects not having the relevant data. Alternatively, the missing data for each subject could be imputed by adding a step for sampling from the full conditional, which is typically easily accomplished. If the data are imputed, the other sampling steps would not need to be

modified. However, we recommend the former approach, as conditioning on imputed data in updating $\{S_i\}$ and the atoms Θ can lead to worse mixing of the MCMC algorithm.

There are two alternative strategies one can take for prediction of Y given $X = x$. Firstly, one could follow a semi-supervised learning approach in which posterior computation is conducted jointly for a sample of labeled subjects $i = 1, \ldots, n_0$ having data $\{x_i, y_i\}$ for both the predictors and response and for a sample of unlabeled subjects $n_0 + 1, \ldots, n$ having data $\{x_i\}$ only for the predictors. This is a special case of the missing data problem and we would recommend imputing $y_i \sim \mathcal{K}^{(y)}(\Theta_{S_i}^{(y)})$ for $i = n_0 + 1, \ldots, n$ but not using these imputed values in updating $\{S_i\}$ and the atoms Θ. Under an appropriate loss function (e.g., squared error for $\mathcal{Y} = \Re$ or 0-1 for $\mathcal{Y} = \{0, 1\}$), one can estimate an optimal predictive value for each subject based on these samples, while also obtaining predictive intervals to accommodate uncertainty. In addition, the predictive density for y_{n+1} given an arbitrary x_{n+1} value can be estimated by averaging expression (8) across MCMC iterations after burn-in for a dense grid of possible values for y_{n+1}. As a second strategy, we could implement the MCMC algorithm for an initial sample of subjects, and avoid rerunning the MCMC algorithm as new unlabeled subjects are obtained. This approach does not utilize information in the predictors for the new subjects in calculating the posterior for the parameters, but may lead to a substantial computational gain in some cases.

3   Some Examples

3.1   Classification from Euclidean Predictors

To highlight differences with the approach of [23] described in Section 1, we initially consider the case in which $y_i \in \mathcal{Y} = \{1, \ldots, c\}$ and $x_i \in \Re^p$, so that $\mathcal{X}_j = \Re$, for $j = 1, \ldots, p$. [23] proposed a joint DPM model for (y_i, x_i). Within each cluster produced by the DP, the joint

distribution of $(y_i, x_i | S_i)$ was characterized as a product of independent normals for x_i with a multinomial logit model for y_i given x_i. Even for a single multinomial logit model, posterior computation can be quite involved, particularly as the number of predictors increases. Hence, posterior computation for the joint DPM of a multinomial logit and a product of normals can be computationally expensive.

Our product kernel mixture approach is considerably simpler when conjugate priors are chosen. In particular, let

$$
\begin{aligned}
y_i &\sim \sum_{l=1}^{c} \psi_{S_i l} \delta_l, \quad \psi_h = (\psi_{h1}, \ldots, \psi_{hc})' \sim \text{Dir}(b_1, \ldots, b_c) \\
x_{ij} &\sim N(\mu_{S_i j}, \sigma_{S_i j}^2), \quad (\mu_{hj}, \sigma_{hj}^{-2}) \sim \text{N-Ga},
\end{aligned}
\tag{14}
$$

where N-Ga denotes a conjugate normal-gamma prior jointly for the mean and precision in each component. Posterior computation is embarrassingly easy for the model following the algorithm of Section 2.1, and noting that step 3 corresponds to sampling from a conjugate Dirichlet and step 5 to sampling from normal-gamma conditional posteriors.

Conditionally on the parameters and mixture weights, the classification function is

$$
\Pr(Y = y \mid X = x, \pi, \Theta) = \sum_{h=1}^{k} \frac{\pi_h \psi_{hy} N_p(x; \mu_h, \Sigma_h)}{\sum_{l=1}^{k} \pi_l \psi_{ly} N_p(x; \mu_l, \Sigma_l)}, \quad y = 1, \ldots, c.
\tag{15}
$$

Hence, the conditional probability of $Y = y$ given predictors $X = x$ is expressed as a convex combination of k basis probability vectors $\psi_h = (\psi_{h1}, \ldots, \psi_{hc})'$, for $h = 1, \ldots, k$. The weights on these probability vectors are proportional to a global weight π_h times a Gaussian kernel that decreases with distance between the individuals' predictors x and the location of the kernel μ_h.

In addition to leading to simple posterior computation, the classification function in (15) is extremely flexible. To illustrate this heuristically, we consider the case in which $c = 2$ and let $\psi_h = \Pr(y_i = 2 \mid S_i = h)$ to simplify notation. Then, $\psi_h \in [0, 1]$ is a probability placed at location $\mu_h \in \Re$, for $h = 1, \ldots, k$, with the conditional probability of $Y = 2$ given

$X = x$ a weighted average of the ψ_h's where the weights are proportional to global weights π_h times normal kernels that decrease with Euclidean distance between x and μ_h. Figure 1 shows realizations of the classification function in a simple toy example in which $k = 2$, $p = 1$, $\pi_1 \sim \text{Unif}(0, 1)$, $\mu_h \sim N(0, 1)$, $\sigma_h^{-2} \sim \text{Ga}(1, 1)$, and $\psi_h \sim \text{Unif}(0, 1)$, for $h = 1, 2$. Even with only two kernels, an amazing variety of curves can be generated, and as the number of kernels increases any smooth classification function can be approximated.

3.2   Classification from Functional Predictors

An appealing aspect of the product kernel mixture approach is that it can be easily adapted to accommodate complex high-dimensional and functional predictors. Essentially, as long as we have a hierarchical model for the response and each of the p predictors, we can implement the approach. We simply specify independent hierarchical models for each predictor and the response and then link them together through the shared cluster index S_i. To illustrate this, suppose that $y_i \in \{0, 1\}$ is an indicator of an adverse response, $x_{i1} \in \Re$ is continuous, $x_{i2} \in \{0, 1\}$ is binary and x_{i3} is a function. For example, [4] considered an application in which the functional predictor is the trajectory in a progesterone metabolite (PdG) in early pregnancy starting at conception and y_i is an indicator of early pregnancy loss. In this application, x_{i1} is age of the woman, while x_{i2} is an indicator of prenatal exposure of the woman to her mother's cigarette smoking. Following a related approach to that described above, we simply let $y_i \sim \text{Bernoulli}(\psi_{S_i})$, $x_{i1} \sim N(\mu_{S_i1}, \sigma_{S_i1}^2)$, and $x_{i2} \sim \text{Bernoulli}(\mu_{S_i2})$. Then, for the functional trajectory x_{i3} data, we specify the following hierarchical model:

$$
\begin{aligned}
w_{ij} &= f_i(t_{ij}) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \\
f_i(t) &= \sum_{d=1}^{q} \beta_{S_i d} b_d(t), \quad \beta_h = (\beta_{h1}, \ldots, \beta_{hq})' \sim P_{03},
\end{aligned}
\tag{16}
$$

where $x_{i3} = (w_{i1}, \ldots, w_{in_i})'$ is a vector of error-prone measurements of PdG for woman i, $t_i = (t_{i1}, \ldots, t_{in_i})'$ are the measurement times, $f_i(t)$ is a smooth trajectory for woman i,
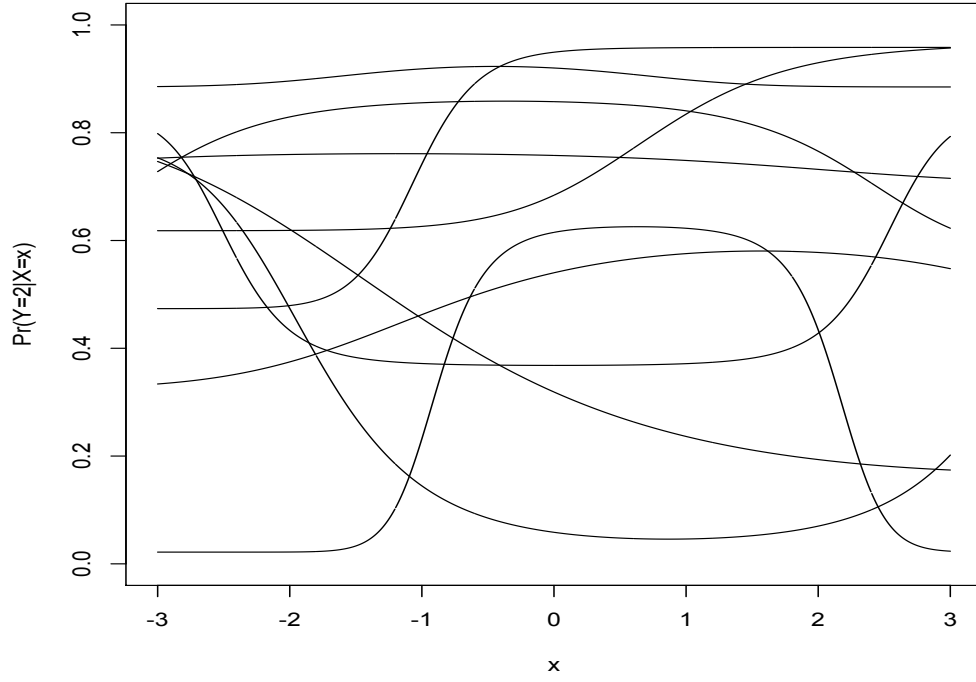
Figure 1. Realizations from prior (15) for a binary classification function in the $k = 2$ and $p = 1$ case.

$\{b_d\}_{d=1}^{q}$ are pre-specified basis functions, and β_h is a vector of basis coefficients specific to cluster h. The base measure P_{03} can include a variable selection structure to allow basis selection to vary across the clusters.

Under this model, cluster h consists of those women having $S_i = h$. This cluster has a distinct probability of early loss, ψ_h, distribution of ages, $N(\mu_{h1}, \sigma_{h1}^2)$, probability of exposure, μ_{h2}, and trajectory in PdG, $f(t) = b(t)'\beta_h$. It is straightforward to apply the MCMC algorithm of Section 3.2 to perform posterior computation, and use the resulting model for prediction of early pregnancy loss based on the woman's predictors and a partial time series of PdG measurements. We could even extend the model to include additional information, such as an ultrasound image in early pregnancy, with an additional hierarchical model specified for that image. This model is an alternative to the discriminant analysis approach described in [7], which instead relies on using a dependent Dirichlet process for the distribution of the function within each pregnancy outcome category.

Although the proposed product kernel mixture approach is clearly highly flexible and easy to implement, one potential concern is that the implicit assumption of global clustering may be overly restrictive in applications involving multivariate and functional predictors. For example, it may be unreasonable to assume that the PdG trajectories in early pregnancy are exactly the same for any two women. In addition, two women in the same cluster may have similar overall profiles in terms of most of their predictors but there may be individual predictors that deviate from this profile. Hence, it may be more realistic to allow local violations of the global clustering assumption in which a subject is still allocated to a global cluster S_i but certain predictors for that subject or parameters for a given predictor are "contaminated" and are instead allocated to a different component. [9] proposed a local partition process (LPP) prior, which allows such contamination, and it is straightforward to modify the above computational algorithm to use a finite approximation to the LPP prior in place of the finite approximation to the Dirichlet process.

In cases in which the response is continuous instead of discrete, we can similarly accommodate mixtures of discrete, continuous and even functional predictors. We simply let $y_i \sim N(\psi_{S_i}, \tau_{S_i})$ in place of the Bernoulli in the above specification. This results in a simple

method for density regression [12], which can easily accommodate a rich variety of predictors. The conditional density of the response will change flexibly with these predictors.

3.3 Classification & Testing from Predictors on a Manifold

Now consider the case in which $x_{i1} \in \mathcal{X}_1$, with $\mathcal{X}_1$ a known non-Euclidean manifold, such as a hypersphere or planar shape space. We consider two applications. The first is to morphometrics in which there is interest in studying the shape of an organism and using shape for classification. In this setting, the data collected for an organism may consist of the (x, y) coordinate locations of landmarks, which correspond to pre-specified features of the organism. For example, [8] consider data on the location of eight landmarks on the midline plane of 2d images of 29 male and 30 female gorilla skulls. For anthropologists, it is of interest to be able to input the landmark locations in a classifier which then predicts gender. In studying shape, it is important to be robust to translations, rotations and scaling of the (x, y) coordinates. Hence, we do not want to specify a model directly for the Euclidean locations of the landmarks but instead want to remove translations, rotations and scaling from the data and build a nonparametric model directly on the planar shape space. The questions that then arise include how to nonparametrically estimate a shape density across organisms, test for differences in shape between groups and obtain a good classifier based on shape features?

Another motivating application is to classification and testing of differences between groups based on features corresponding to locations on a hypersphere. One example is to global data on volcanoes. For volcano i $(i = 1, \ldots, n)$, $y_i \in \{1, \ldots, c\}$ denotes the volcano type and x_i denotes the location on the globe. Spatial data often have a similar structure in which observations are collected at different locations on the globe, with observation locations potentially informative about the distribution of "marks" at each location. Although spatial

data on the earth are often treated as Euclidean, this can lead to substantial artifacts in large scale spatial data in which information is available not only for a small location region of the earth but also for a wide area. Treating the locations as Euclidean distorts the geometry and does not account for the fact that locations that are very far apart in Euclidean distance can be close together when considering the geodesic distance on the sphere.

In both the gorilla skull shape and volcano location applications, to apply the general product kernel mixtures methodology of Section 2, we require a kernel mixture model for density estimation on a compact Riemannian manifold. [2] proposed a general class of kernel mixture models for Bayesian density estimation on manifolds, providing sufficient conditions for Kullback-Leibler support of the prior and weak posterior consistency in density estimation. It is important to show that the prior has large support, because even if a prior seems flexible it may rule out many *a priori* plausible models. [2] considered Dirichlet process mixtures of complex Watson kernels for planar shape data and Dirichlet process mixtures of von Mises kernels for hyperspherical data, showing in both cases that the kernels satisfy the sufficient conditions for large prior support and weak consistency. [3] further developed the theory in providing sufficient conditions for strong consistency in density estimation on compact metric spaces including manifolds. Complex Watson and von Mises kernels are computationally convenient choices in having conjugacy properties making implementation of the MCMC algorithm of Section 2.2 and related algorithms straightforward.

[1] considered the special case of the product kernel mixture model (4) - (5) in which $\mathcal{Y} = \{1, \ldots, c\}$, $p = 1$, $\mathcal{X}_1$ corresponds to a compact Riemannian manifold, and a Dirichlet process prior is assumed for the mixing measure. In this case, [1] developed theory giving conditions for strong consistency in estimating the classification function $\Pr(Y = y \mid X = x)$. This implies that regardless of the true relationship between each class probability and the predictors lying on a manifold, the posterior for the classification function will concentrate exponentially fast around the truth increasingly as the sample size increases. This class of

models is appropriate for both the gorilla skull shapes and volcano locations applications. In ongoing work, it will be interesting to generalize the theory beyond the classification setting to include arbitrary product kernel mixtures for any combination of data types.
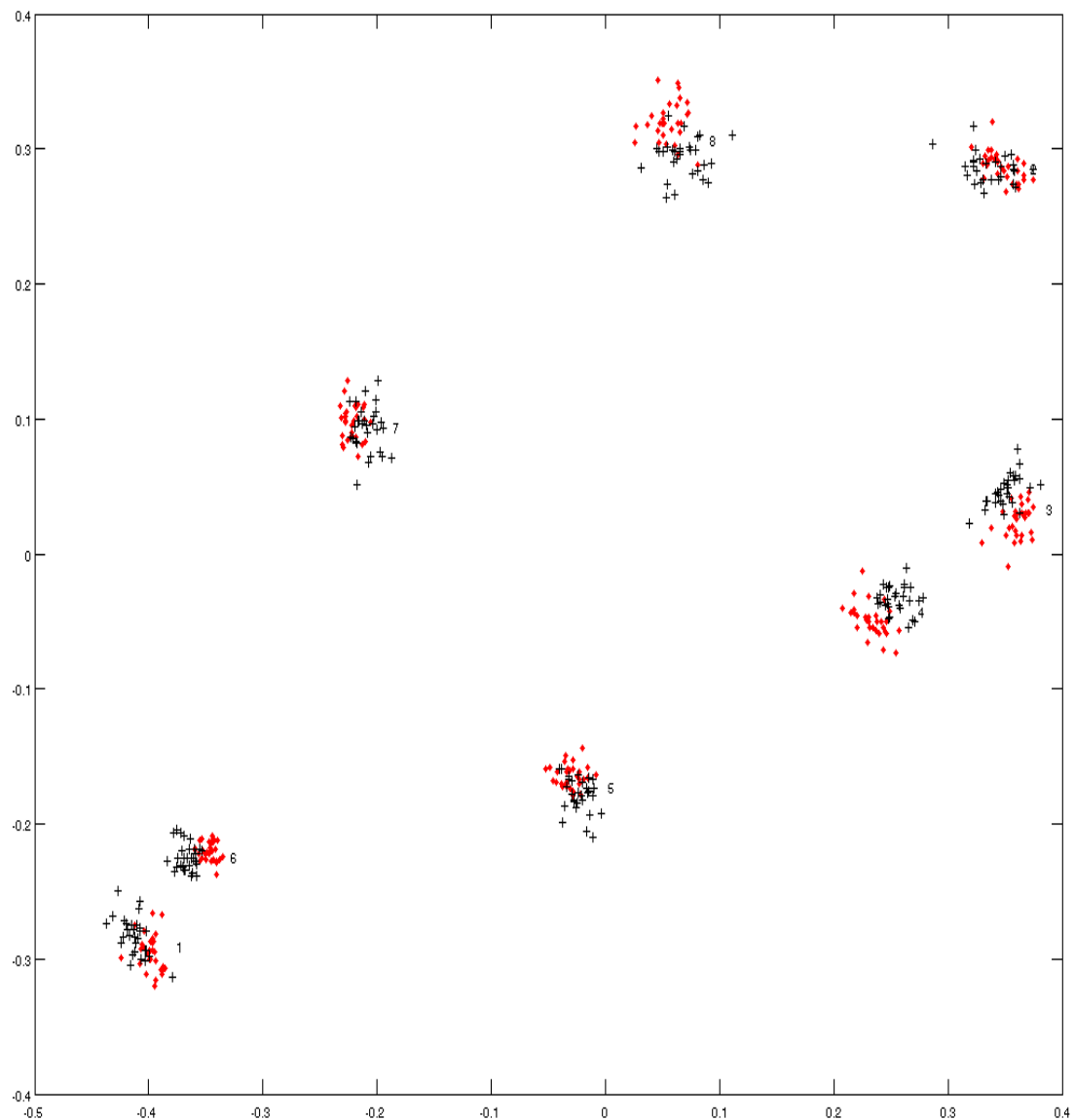

Figure 2. Preshapes for 29 male and 30 female gorilla skulls based on data for 8 landmarks.

Focusing on the gorilla skull shape application, let $z = (z_1, \ldots, z_k)' \in C^k$ denote the

complex k-ad vector of landmark locations, let $z_c = z - \bar{z}$ denote the centered k-ad obtained by subtracting the centroid, and let w denote the preshape corresponding to a point on the complex sphere obtained by normalizing z_c to removing scaling. The similarity shape of z is the orbit of w under all rotations in 2D, with the space of all such orbits corresponding to the planar shape space Σ_2^k. For gorilla i, $y_i \in \{0, 1\}$ with $y_i = 1$ for males and $y_i = 0$ for females, and let $x_i \in \mathcal{X} = \Sigma_2^k$ denote the similarity shape of the skull. The gorilla skull preshapes for females and males are shown in Figure 2. To complete a specification of the model, we let $y_i \sim \text{Bernoulli}(\psi_{S_i})$ and $x_i \sim \text{CW}(\mu_{S_i}, \kappa)$, where the complex Watson distribution corresponds to

$$\text{CW}(m; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa |z^* v|^2),$$

where z, v are preshapes of $m, \mu \in \Sigma_2^k$, $*$ denotes the complex conjugate transpose, μ is the extrinsic mean, κ is a measure of concentration, and $c(\kappa)$ is a normalizing constant. Posterior computation is straightforward applying a related algorithm to that described in Section 2.1 and we obtain good performance in out-of-sample classification. [1] modified this approach to test for differences in the shape distributions between males and females, while showing Bayes factor consistency under the alternative hypothesis.

Considering the volcano locations application, we let $y_i \sim \sum_{l=1}^{3} \psi_{S_i y} \delta_y$ and $x_i \sim \text{vMF}(\mu_{S_i}, \kappa)$, with the kernel for the volcano locations corresponding to the von Mises-Fisher distribution,

$$\text{vMF}(x; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa x' \mu),$$

where μ is the extrinsic mean and κ is a measure of concentration. We focused on the $n = 999$ volcanoes of the three most common types, including strato, shield and submarine. Again, a simple Gibbs sampler can be implemented following the algorithm of Section 2.2. Based on this, we obtain better out-of-sample performance in classifying the volcano types than in using discriminant analysis based on mixtures of Gaussians. Applying the [1] testing

approach, there is strong evidence that the varying types of volcanoes have differing spatial distributions.

4 Discussion

This article has proposed a simple and very general strategy for flexible joint modeling of data having a variety of supports via a discrete mixture of product kernels. The emphasis has been on nonparametric modeling of the conditional distribution of response data Y given predictors $X = \{X_1, \ldots, X_p\}$, but there is no need to specify a response when interest is in modeling dependence. The framework can accommodate joint modeling of a rich variety of data structures such as functions, images, shapes, data with support on a manifold and mixed discrete and continuous vectors. As long as we can specify parametric hierarchical models for the different component data, then we can build a joint model through linking the component models through a shared cluster index. If MCMC algorithms are available for posterior computation in the separate component models, then these algorithms can be trivially adapted to accommodate joint modeling under the proposed framework. Although the model seems overly simple, in many cases under weak restrictions on the kernels and true data-generating model, one can obtain full support and weak and strong posterior consistency in estimating the joint distribution and conditionals given predictors. [1] showed this in a particular case, but the theory can conceptually be generalized.

Along with the positive characteristics of this approach come some possible concerns and limitations. Firstly, considering the conditional modeling case, the model explicitly treats the predictors as random, which may not be a realistic representation of reality in certain cases, such as when predictors are fixed by design. That said, in many cases predictors that are treated as fixed may be more realistically modeled as random and many methods for accommodating missing predictors treat predictors as random in performing imputation. If

predictors are truly fixed, then one can potentially view the X likelihood as an auxiliary model that is just incorporated to induce a coherent and flexible model for the conditional of Y given X. This view was advocated in [20].

A potentially more serious concern is somewhat subtle. The argument is as follows. Suppose we are interested in the conditional distribution of Y given X and have no interest in inferences on the marginal of X. The proposed discrete mixture of product kernels model nonetheless models the joint of Y and X. In updating the prior with the likelihood of the data, the intrinsic Bayes penalty for model complexity will tend to lead to a parsimonious characterization of the data, with parsimony manifest in discrete mixture models partly through allocation of the n subjects to $k_n << n$ clusters. The posterior on the allocation to clusters is driven by a "desire" of the Bayesian invisible hand to allocate clusters in such a way as to obtain a high marginal likelihood with relatively few clusters occupied. In certain cases, such as when there are many predictors or more information in the predictor component of the likelihood, the marginal of X can dominate and play much more of a role in allocation to clusters. This can lead to poor performance in estimating the conditional of Y given X and in predicting Y given X. Even when the conditional likelihood of Y given X has an important impact on clustering, there can be extra clusters introduced just to better fit the marginal of X even though these clusters may just degrade the performance in prediction. A potential fix-up to these issues is to include separate but dependent cluster indices for the predictor and response component [10].

An alternative, which has a number of advantages, is to avoid modeling the joint of Y and X and to instead define a model directly for the conditional of Y given X. There is an increasing literature on such conditional modeling approaches [12, 11, 15, 6, 25], though they remain to be developed for general predictors X, including shapes and predictors with support on a variety of manifolds. We plan to pursue this and to develop theory of large support, posterior consistency and rates of convergence in ongoing work.

References

[1] A. Bhattacharya and D. Dunson. Nonparametric Bayes classification and testing on manifolds with applications on hypersphere. *Discussion Paper Duke University*, 2010.

[2] A. Bhattacharya and D. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 2010. To Appear.

[3] A. Bhattacharya and D. Dunson. Strong consistency of nonparametric Bayes density estimation on compact metric spaces. *Discussion Paper Duke University*, 2010.

[4] J. Bigelow and D. Dunson. Bayesian semiparametric joint models for functional predictors. *J. Am. Statist. Ass.*, 104:26–36, 2009.

[5] D. Blackwell and J.B. MacQueen. Ferguson distributions via Polya urn schemes. *Annals of Statistics*, pages 353–355, 1973.

[6] Y. Chung and D.B. Dunson. Nonparametric Bayes conditional distribution modeling with variable selection. *J. Am. Statist. Ass.*, 104:1646–1660, 2009.

[7] R. De la Cruz-Mesia, F.A. Quintana, and P. Müller. Semiparametric bayesian classification with longitudinal markers. *Applied Statist.*, 56:119–137, 2007.

[8] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. Wiley N.Y., 1998.

[9] D.B. Dunson. Nonparametric bayes local partition models for random effects. *Biometrika*, 96:249–262, 2009.

[10] D.B. Dunson, A.H. Herring, and A.M. Siega-Riz. Bayesian inferences on changes in response densities over predictor clusters. *Journal of the American Statistical Association*, 103:1508–1517, 2008.

[11] D.B. Dunson and J-H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2009.

[12] D.B. Dunson, N. Pillai, and J-H. Park. Bayesian density regression. *Journal of the Royal Statistical Society B*, 69:163–183, 2007.

[13] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[14] T. S. Ferguson. Prior distributions on spaces of probability measures. *Ann. Statist.*, 2:615–629, 1974.

[15] R. Fuentes-Garcia, R.H. Mena, and S.G Walker. A nonparametric dependent process for bayesian regression. *Statistics and Probability Letters*, 79:1112–1119, 2009.

[16] H. Ishwaran and M. Zarepour. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12:941–963, 2002.

[17] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[18] M. Kali, J.E. Griffin, and S.G. Walker. Slice sampling mixture models. *Statistics and Computing*, 2009. Online.

[19] P. Müller, A. Erkanli, and M. West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83:67–79, 1996.

[20] J-H. Park and D.B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20:1203–1226, 2010.

[21] G. Roberts and J. Rosenthal. Coupling and ergodicity of adaptive mcmc. *Journal of Applied Probability*, 44:458–475, 2007.

[22] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in over-fitted mixture models. *Technical Report*, 2010.

[23] B. Shahbaba and R. Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

[24] M. Stephens. Dealing with label switching in mixture models. *J. R. Statist. Soc. B*, 62:795–809, 2000.

[25] S.T. Tokdar, Y.M. Zhu, and J.K. Ghosh. Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Analysis*, 5:319–344, 2010.

[26] C. Yau, O. Papaspiliopoulos, G.O. Roberts, and C. Holmes. Nonparametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes. *J. R. Statist. Soc. B*, 2010. To Appear.