Classification via Bayesian Nonparametric Learning of Affine Subspaces

Garritt Page
Departamento de Estadística
Pontificia Universidad Católica de Chile
page@mat.puc.cl

Abhishek Bhattacharya
Indian Statistical Institute
Kolkata India
abhishek@isical.ac.in

David Dunson
Department of Statistical Science
Duke University
dunson@stat.duke.edu

December 12, 2012

Abstract

It has become common for data sets to contain large numbers of variables in studies conducted in areas such as genetics, machine vision, image analysis and many others. When analyzing such data, parametric models are often too inflexible while nonparametric procedures tend to be non-robust because of insufficient data on these high dimensional spaces. This is particularly true when interest lies in building efficient classifiers in the presence of many predictor variables. When dealing with these types of data, it is often the case that most of the variability tends to lie along a few directions, or more generally along a much smaller dimensional submanifold of the data space. In this article, we propose a class of models that flexibly learn about this submanifold while simultaneously performing dimension reduction in classification. This methodology, allows the cell probabilities to vary nonparametrically based on a few coordinates expressed as linear combinations of the predictors. Also, as opposed to many black-box methods for dimensionality reduction, the proposed model is appealing in having clearly interpretable and identifiable parameters which provide insight into which predictors are important in determining accurate classification boundaries. Gibbs sampling methods are developed for posterior computation, and the methods are illustrated using simulated and real data applications.

Key Words: Classifier; Dimension reduction; Variable selection; Nonparametric Bayes

1 Introduction

Experiments or studies carried out in areas such as epidemiology, image analysis, and machine vision (to name a few) are producing data sets whose dimension continues to increase. The increased dimension is often caused by collecting a large number of predictor variables with the goal of building efficient classifiers (or regression models) that further understanding regarding the associations between the response of interest (Y) and predictors ($\boldsymbol{X}$). Because such data sets have become commonplace, designing data efficient inference techniques that scale to high dimensional Euclidean and even non-Euclidean spaces have attracted considerable attention in the statistical and machine learning literature.

In addition to being able to scale to higher dimensions, it is often highly desirable for methods to provide insight regarding the underlying mechanisms of the phenomena being studied. For example, one might want to characterize the joint effect of a subset of covariates on an outcome of interest and determine which are important. As an illustration, we consider a sub-study of the US Collaborative Perinatal Project (CPP) (a reproductive epidemiology study) which collected data on pregnancy outcomes together with demographic factors, and levels of exposure to a wide variety of environmental contaminants. Identifying a collection of these variables that influence preterm and/or small-for-gestational-age babies at birth was of particular interest. Also of interest was to ability to create an index of an individual's exposure burden and characterize the joint health effects of the exposures. In situations like these, many methods proposed in the literature (such as support vector machines (SVM), Cortes and Vapnik 1995, neural networks Hastie et al. 2008, and fully Bayesian hierachical probability models Chen et al. 2010) are inadequate as they are algorithmic or highly parameterized black boxes and apart from classification (or curve fitting), provide no further information specific to the problem being studied. We will revisit this study in Section 5.3.

When dealing with a high dimensional feature space, it is typically the case that parametric models are too rigid and don't adapt well to high dimensions while flexible nonparametric approaches suffer from the well known curse of dimensionality. With this in mind, an appealing approach is to make procedures more scalable to high dimensions by learning a lower dimensional subspace the covariates are concentrated near. If the subspace were known, then one could model the projections of $\boldsymbol{X}$ onto that subspace with a nonparametric density model, while using some simple parametric distribution on the orthogonal residual vector. A robust classifier (or regression model) would be attained by fitting a flexible model on only a selected few coordinates of the projections. These coordinates would then act as surrogate predictors that more efficiently explain the variability in Y.

Building classifiers using coordinates of projections is similar to what has been called sufficient dimension

reduction (SDR) (Zhu and Zeng (2006), Li (1991) and Cook and Weisberg (1991)). Typically, this is a two stage approach the first involving the estimation of a lower dimensional subspace (often called the central subspace) and the second using these coordinates as predictors in some classification (or regression) method (Li and Wang 2007). To avoid the two stage approach (and hence incorporate subspace estimation uncertainty in inferences) Wang and Xia (2008) proposed an SDR approach that handles subspace estimation and regression simultaneously by modeling the conditional density of Y given sufficient predictors using kernel smoothing. Recently Reich et al. (2011) developed a Bayesian method that simultaneously performs regression while estimating the central subspace by placing a Gaussian prior on a basis of the subspace and using a finite mixture to model $Y|\boldsymbol{X}$. Tokdar et al. (2010) also developed a Bayesian approach but assign a uniform prior to the space of d-dimensional linear subspaces of $\Re^p$ (where d is dimension of subspace and p dimension of feature space) and modeled the density of $Y|\boldsymbol{X}$ with log Gaussian process.

We approach the problem from a completely different perspective. Instead of modeling $Y|\boldsymbol{X}$ directly, we first propose modeling $(Y, \boldsymbol{X})$ jointly through a mixture of product kernels and then employ the flexible conditional model that is induced by the joint. The idea of using a joint model to induce a flexible model on the conditional was first proposed by Müller et al. (1996) and developed further by Bhattacharya and Dunson (2012) and Dunson and Bhattacharya (2011). However, this earlier work lacked the dimensionality reduction component. Our proposed framework very flexibly and uniquely identifies a lower dimensional affine subspace, while simultaneously modeling the coordinates of the projections of $\boldsymbol{X}$ onto that subspace using an infinite mixture of Gaussians. Then the data component orthogonal to that subspace is modeled independently with a zero mean Gaussian.

Among all possible coordinate choices, we use isometric coordinates (those which preserve the geometry of the space). To obtain such coordinates, only orthogonal bases for the subspace will be considered. In addition to interpretability and identifiability, there are computational advantages as matrix inversion is equal to transpose. Thus mixture component contours are not required to be homogeneous, but rather can take on a sparse singular value decomposition type representation. This accommodates potentially large non-homogeneous covariance matrices without sacrificing flexibility. An important feature of the methodology is that intuitively appealing interpretations accompany model parameters and therefore provide information regarding covariates (or linear combinations thereof) that directly influence the response. Such interpretability is crucial in many applications, such as epidemiology, and there is a clear lack of interpretability for most methods for flexible classification or prediction from correlated covariates.

The remainder of this article is organized as follows. Section 2 provides a few required geometric ideas

and introduces the notation needed to describe the model. Section 3 first details the marginal model for $\boldsymbol{X}$ and then the classification model. Additionally, theoretical results dealing with posterior consistency, model identifiability and parameter estimation are provided. Section 4 details computational strategies while Section 5 provides some numerical examples using simulated and real data. We finish with some concluding remarks in Section 6.

2 Preliminaries

Before detailing general modeling strategies, we define a few necessary terms. Additionally, since we adopt a geometric perspective, a brief background to some basic geometric ideas is provided.

A k-dimensional affine subspace on $\Re^m$ can be expressed as $S = \{\boldsymbol{R}\boldsymbol{y} + \boldsymbol{\theta} \colon \boldsymbol{y} \in \Re^m\}$ where $\boldsymbol{R}$ is a $m \times m$ rank k *projection matrix* (it satisfies $\boldsymbol{R} = \boldsymbol{R}' = \boldsymbol{R}^2$, $\text{rank}(\boldsymbol{R}) = k$) and $\boldsymbol{\theta} \in \Re^m$ satisfies $\boldsymbol{R}\boldsymbol{\theta} = \boldsymbol{0}$. Since there is a one to one correspondence between the subspace S and the pair $(\boldsymbol{R}, \boldsymbol{\theta})$, identifying the pair $(\boldsymbol{R}, \boldsymbol{\theta})$ is sufficient to learn S. The projection of some $\boldsymbol{x} \in \Re^m$ into S is the point $\boldsymbol{x}_0 \in \Re^m$ that satisfies $\|\boldsymbol{x} - \boldsymbol{x}_0\| = \min\{\|\boldsymbol{x} - \boldsymbol{y}\| : \boldsymbol{y} \in S\}$ (here $\|\cdot\|$ denotes the Euclidean norm). It turns out that for any S defined as above the solution of this minimization problem is $\boldsymbol{x}_0 = \boldsymbol{R}\boldsymbol{x} + \boldsymbol{\theta}$. Therefore, $\boldsymbol{\theta}$ is the projection of the origin into S and $\boldsymbol{R}$ is the projection matrix of the shifted linear subspace $L = S - \boldsymbol{\theta} = \{\boldsymbol{R}\boldsymbol{y} \colon \boldsymbol{y} \in \Re^m\}$. We use $Pr_S(\boldsymbol{x})$ to denote the projection of $\boldsymbol{x} \in \Re^m$ into S.

Let $\boldsymbol{U}$ be a $m \times k$ matrix whose columns $\{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k\}$ form a basis for the column space of $\boldsymbol{R}$. Any $\boldsymbol{x} \in S$ can be given coordinates $\tilde{\boldsymbol{x}} \in \Re^k$ such that $\boldsymbol{x} = \boldsymbol{U}\tilde{\boldsymbol{x}} + \boldsymbol{\theta}$. If $\boldsymbol{U}$ is chosen to be orthonormal (i.e., $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_k$ and $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{U}'$), then the coordinates are *isometric*. That is, they preserve the inner product on S (and hence volume and distances). With such a basis, the projection $Pr_S(\boldsymbol{x})$ of an arbitrary $\boldsymbol{x} \in \Re^m$ into S has isometric coordinates $\tilde{\boldsymbol{x}} = \boldsymbol{U}'\boldsymbol{x}$. Thus, $\boldsymbol{U}$ gives k mutually perpendicular 'directions' to S while $\boldsymbol{\theta}$ may be viewed as the 'origin' of S. We will call $\boldsymbol{\theta}$ the *origin* and $\boldsymbol{U}$ an *orientation* for S.

The *residual* of $\boldsymbol{x} \in \Re^m$ (which we denote as $R_S(\boldsymbol{x}) = \boldsymbol{x} - Pr_S(\boldsymbol{x})$) lies on a linear subspace that is perpendicular to L. That is, $R_S(\boldsymbol{x}) \in S^\perp$ where $S^\perp = \{(\boldsymbol{I} - \boldsymbol{R})\boldsymbol{y} \colon \boldsymbol{y} \in \Re^m\}$. Notice that the projection matrix of $S^\perp$ is $\boldsymbol{I} - \boldsymbol{R}$. Now if we let $\boldsymbol{V}$ denote an orthonormal basis for the column space of $\boldsymbol{I} - \boldsymbol{R}$ (i.e., $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_{m-k}$, $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I} - \boldsymbol{R}$), then isometric residual coordinates are given by $\boldsymbol{V}'\boldsymbol{x} \in \Re^{m-k}$.

If a sample $\boldsymbol{x} \in \Re^m$ lies close to a much smaller dimensional subspace (S), of dimension k, it would be natural to assume that the data residuals are centered around 0 with small variability while the data projected into S comes from a possibly multi-modal distribution supported on S. Figure 1 illustrates such

a sample data cloud. The observations are drawn from a two-component mixture of bivariate normals with cluster centers $(1, 0)$ and $(0, 1)$ and band-width of 0.5. As a result they are clustered around the line $x_1 + x_2 = 1$ which can be characterized as an affine subspace of $\Re^2$ (details follow).

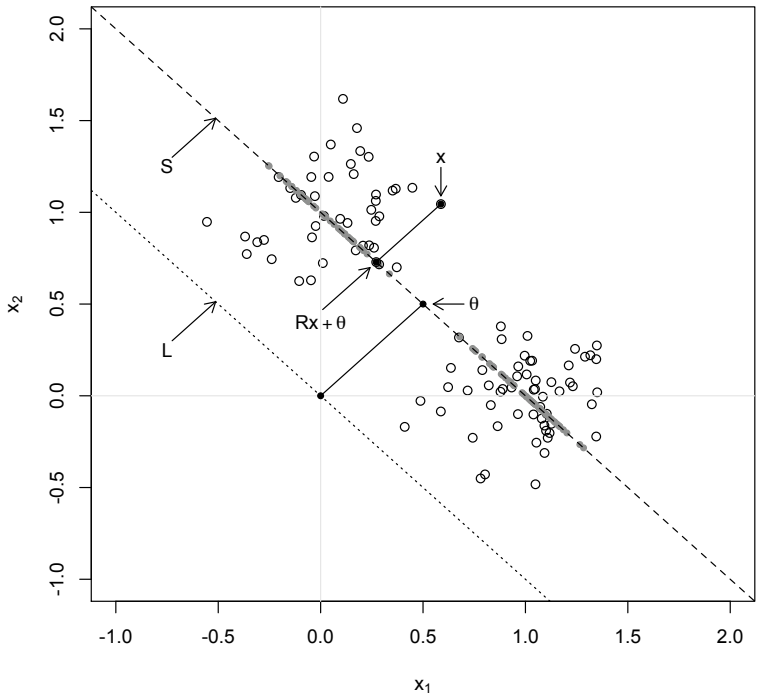

Figure 1: Graphical representation of the affine subspace (S), the orthogonal shift ($\boldsymbol{\theta}$), and the projection of a point into S (these are the solid dots with particular emphasis given to $\boldsymbol{Rx} + \boldsymbol{\theta}$).

Now if Q is a distribution on $\Re^m$ with finite second order moments, then for $d \leq m$ we define the d *principal affine subspace* of Q as

$$\underset{S}{\operatorname{argmin}}\ Risk(S) = \int_{\Re^m} \|\boldsymbol{x} - Pr_S(\boldsymbol{x})\|^2 Q(d\boldsymbol{x}). \tag{2.1}$$

The minimization is carried out over all d-dimensional affine subspaces S. It can be shown that the minimum value always exists and is $\sum_{d+1}^{m} \lambda_j$, where $\lambda_1 \geq \ldots \geq \lambda_m$ are the ordered eigenvalues of Q's covariance matrix. A unique minimizer (which is denoted by S_o) exists if and only if $\lambda_d > \lambda_{d+1}$. When S_o exists, its projection matrix is $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{U}'$ where $\boldsymbol{U}$ is any orthonormal basis for the subspace spanned by a set of d independent eigenvectors that correspond to the first d eigenvalues of the covariance matrix of Q. Furthermore, for $\boldsymbol{\mu}$ the

mean associated with Q, the origin of S_o is $\boldsymbol{\theta} = (\boldsymbol{I} - \boldsymbol{R})\boldsymbol{\mu}$. One thing to notice is that when $d = 0$, then S_o is the point set $\boldsymbol{\mu}$.

In the case that d is unknown, a reasonable strategy to find an optimal value of d would be to minimize the following risk function

$$Risk(d, S) = f(d) + \int_{\Re^m} \|\boldsymbol{x} - Pr_S(\boldsymbol{x})\|^2 Q(d\boldsymbol{x}),\ 0 \leq d \leq m \tag{2.2}$$

in terms of d and S where f is a fixed increasing convex function. If f is linear, say, $f(d) = ad$, $a > 0$, then (2.2) has a unique minimizer if and only if $\lambda_{d+1} < a < \lambda_d$ for some d, with $\lambda_0 = \infty$ and $\lambda_{m+1} = 0$. The minimizing dimension d_o is that value of d while the optimal space S_o is the d_o principal affine subspace. We will call d_o the *principal dimension* of Q. For the observations in Figure 1, the principal dimension is $d_o = 1$ with principal affine subspace

$$S_o = \left\{ \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} \boldsymbol{x} + \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} : \boldsymbol{x} \in \Re^2 \right\}.$$

We end this section by introducing the notation used throughout the remainder of the article. Let $\mathcal{M}(S)$ denote the space of all probabilities on the space S. $M(m, k)$ will denote real matrices of order $m \times k$ (with $M(m)$ denoting the special case of $m = k$), $M^+(m)$ will denote the space of all $m \times m$ positive definite (p.d.) matrices. For $\boldsymbol{U} \in M(m, k)$, $\mathcal{N}(\boldsymbol{U})$ will denote the subspace spanned by the vectors orthogonal to the columns of $\boldsymbol{U}$. We will represent the space of all $m \times m$ rank k projection matrices by $P_{k,m}$. That is,

$$P_{k,m} = \{\boldsymbol{R} \in M(m) \colon \boldsymbol{R} = \boldsymbol{R}' = \boldsymbol{R}^2, \mathrm{rank}(\boldsymbol{R}) = k\}.$$

One important manifold referred to in this paper is the Steifel manifold (denoted by $V_{k,m}$) which is the space whose points are k-frames in $\Re^m$ (here k-frame refers to a set of k orthonormal vectors in $\Re^m$). That is,

$$V_{k,m} = \{\boldsymbol{A} \in M(m, k) : \boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_k\}.$$

The space $V_{k,m}$ is a compact non-Euclidean Riemannian manifold. Because $V_{k,m}$ is embedded in the Euclidean space $M(m, k)$, it inherits the Riemannian inner product which can be used to define the volume form, which in turn can be used as the base measure to construct parametric families of densities. Several parametric densities have been studied on this space, and exact or MCMC sampling procedures exist. For

details, see Chikuse (2003) and Hoff (2007). A density that will be used quite extensively is the so called Bingham-von Mises-Fisher density which has the following form

$$BMF(\boldsymbol{x}; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \propto \text{etr}(\boldsymbol{A}'\boldsymbol{x} + \boldsymbol{C}\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}).$$

The parameters are $\boldsymbol{A} \in M(k, m)$, $\boldsymbol{B} \in M(k)$ symmetric and $\boldsymbol{C} \in M(m)$, while etr denotes exponential trace. As a special case, we obtain the uniform distribution which has the constant density $1/\text{Vol}(V_{k,m})$.

3 Principal Subspace Classifier

The principal aim of this article is to develop *interpretable* methods for efficiently modeling the conditional of Y given $\boldsymbol{X}$, that can be used in classification, conditional density estimation or prediction. Since we will approach this problem by modeling $(Y, \boldsymbol{X})$ jointly with a mixture of product kernels, we first focus on the model for $\boldsymbol{X}$ (which is a novel contribution of this article). The model is developed from a geometric perspective focusing on the density estimation of $\boldsymbol{X}$. Theory regarding the method's soundness, parameter identifiability, and parameter estimation is also provided.

3.1 Density Estimation of $\boldsymbol{X}$ via Coordinate Modeling

Consider a random variable $\boldsymbol{X} \in \Re^m$. Let there be a k dimensional affine subspace S, $0 \leq k \leq m$, with projection matrix $\boldsymbol{R}$ and origin $\boldsymbol{\theta}$ such that the projection of $\boldsymbol{X}$ into this subspace follows a location mixture density on the subspace (with respect to its volume form) given by

$$\boldsymbol{z} = Pr_S(\boldsymbol{X}) \sim f(\boldsymbol{z}|\boldsymbol{U}, \boldsymbol{A}) = \int_S (2\pi)^{-k/2} |\boldsymbol{U}'\boldsymbol{A}\boldsymbol{U}|^{1/2} \exp\left\{ -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{w})'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{w}) \right\} Q(d\boldsymbol{w})$$

where $\boldsymbol{z} \in S$, $Q \in \mathcal{M}(S)$, $\boldsymbol{U} \in V_{k,m}$ is any orientation for S, and $\boldsymbol{A}$ is a $m \times m$ positive semi-definite (p.s.d.) matrix such that $\boldsymbol{U}'\boldsymbol{A}\boldsymbol{U} \in M^+(k)$ is positive definite (p.d.). Note that the density expression depends on $\boldsymbol{U}$ only through $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{U}'$.

In other words, conditional on latent variable $\boldsymbol{w} \in S$, the projection follows a Gaussian density conditioned on subspace S: $f(\boldsymbol{z}|\boldsymbol{w}, \boldsymbol{U}, \boldsymbol{A}) \propto \exp\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{w})'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{w})\}$ and we integrate out $\boldsymbol{w}$ to obtain $f(\boldsymbol{z}|\boldsymbol{U}, \boldsymbol{A})$. A general choice of $\boldsymbol{A}$ for this to be a valid density (besides being p.d.) could be $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{U}'$ with $\boldsymbol{\Sigma}_0 \in M^+(k)$. Using change of variables it can be shown that the isometric coordinates $\boldsymbol{U}'\boldsymbol{X}$ of $Pr_S(\boldsymbol{X})$

follow a non-parametric Gaussian mixture model on $\Re^k$ given by

$$\boldsymbol{U}'\boldsymbol{X} \sim \int_{\Re^k} N_k(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}_0) P(d\boldsymbol{\mu}),\ P \in \mathcal{M}(\Re^k), \tag{3.1}$$

where $\boldsymbol{\mu} = \boldsymbol{U}'\boldsymbol{w}$ for $\boldsymbol{w} \in S$.

Independently, we assume that the residual $R_S(\boldsymbol{X})$ follows a mean zero homogeneous Gaussian density (with respect to its volume form) conditioned on $S^\perp$ given by

$$\boldsymbol{r} = R_S(\boldsymbol{X}) \sim g(\boldsymbol{r}|\sigma) \propto \sigma^{-(m-k)} \exp\left\{ -\frac{\|\boldsymbol{r}\|^2}{2\sigma^2} \right\},$$

$\boldsymbol{r} \in S^\perp$ and parameter $\sigma > 0$. If $k = m$, then $S^\perp = \{0\}$ and $R_S(\boldsymbol{X}) = 0$. As a result, with any orientation $\boldsymbol{V} \in V_{m-k,m}$ for $S^\perp$, the isometric coordinates $\boldsymbol{V}'\boldsymbol{X}$ of $R_S(\boldsymbol{X})$ follow the Gaussian density

$$\boldsymbol{V}'\boldsymbol{X} \sim N_{m-k}(\boldsymbol{V}'\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_{m-k}). \tag{3.2}$$

Combine equations (3.1) and (3.2) to get the full density of $\boldsymbol{X}$ as

$$\boldsymbol{X} \sim f(\boldsymbol{x}; \boldsymbol{\Theta}) = \int_{\Re^k} N_m(\boldsymbol{x}; \phi(\boldsymbol{\mu}), \boldsymbol{\Sigma}) P(d\boldsymbol{\mu}), \tag{3.3}$$

$$\phi(\boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{\mu} + \boldsymbol{\theta},\ \boldsymbol{\Sigma} = \boldsymbol{U}(\boldsymbol{\Sigma}_0 - \sigma^2 \boldsymbol{I}_k)\boldsymbol{U}' + \sigma^2 \boldsymbol{I}_m, \tag{3.4}$$

with parameters $\boldsymbol{\Theta} = (k, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\Sigma}_0, \sigma, P)$. We will refer to (3.3) and (3.4) as the Principal Subspace Density (PSD) model.

The final form of $\boldsymbol{X}$'s density model is attractive. We are able to use a flexible multimodal density model on a few data coordinates (which are chosen using a suitable basis) and an independent centered Gaussian structure on the remaining coordinates allows efficient density estimation on potentially high dimensional spaces. In some sense what has been developed here from a geometric perspective could be considered a Bayesian nonparametric extension of the probabilistic principal component analysis of Tipping and Bishop (1999) and Nyamundanda et al. (2010). Furthermore, the model could also be thought of as a nonparametric extension of the Bayesian Gaussian process latent variable models of Titsias and Lawrence (2010) and SVD models of Hoff (2007). A few more comments regarding the density model follow.

An alternative way to identify the intercept $\boldsymbol{\theta}$ would be to set it equal to $\mathrm{E}(\boldsymbol{X})$. However, this would require the prior on P to be such that $\bar{\boldsymbol{\mu}} \equiv \int \boldsymbol{\mu} P(d\boldsymbol{\mu}) = 0$ making the commonly used Dirichlet process an

inappropriate prior for P. For this reason, we set $\boldsymbol{\theta}$ to be the origin of S instead.

With $\boldsymbol{\Sigma}_0$ p.d. and $\sigma^2 > 0$, the within cluster covariance $\boldsymbol{\Sigma}$ lies in $M^+(m)$ and has a sparse representation without being homogeneous. The residual variance σ^2 dictates how "close" $\boldsymbol{X}$ lies to S, with $\sigma^2 = 0$ implying that $\boldsymbol{X} \in S$. In (3.3), one may mix across $\boldsymbol{\Sigma}_0$ by replacing $P(d\boldsymbol{\mu})$ by $P(d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}_0)$ and achieve more generality.

Without loss of generality, we can make model (3.3) even more sparse by allowing $\boldsymbol{\Sigma}_0$ to be a p.d. diagonal matrix. To show that no generality is lost, consider a singular value decomposition (s.v.d.) of an unstructured $\boldsymbol{ODO}' = \boldsymbol{\Sigma}_0$, with $\boldsymbol{O} \in O(k)$, and replace $\boldsymbol{\Sigma}_0$ by diagonal $\boldsymbol{D}$, and $\boldsymbol{U}$ by $\boldsymbol{UO}'$. If P is appropriately transformed, then the model is unaffected. With a diagonal $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}$ has k eigenvalues from $\boldsymbol{\Sigma}_0$ and the rest all equal to σ^2. Furthermore, the columns of $\boldsymbol{U}$ are the orthonormal eigenvectors corresponding to $\boldsymbol{\Sigma}_0$.

It is straightforward to check that $S = S_o$ for the model if and only if $\boldsymbol{\Sigma}_0 + \int_{\Re^k} (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})' P(d\boldsymbol{\mu}) > \sigma^2 I_k$. (Here $\boldsymbol{A} > \boldsymbol{B}$ refers to $\boldsymbol{A} - \boldsymbol{B}$ being p.d.) This holds, for example, when $\boldsymbol{\Sigma}_0 \geq \sigma^2 I_k$ and P is non-degenerate. Further, k is the principal dimension of $\boldsymbol{X}$ for a range of risk functions as in (2.2) with linear f.

3.2 Consistency of Posterior Distribution

When working with nonparametric Bayesian models it is important to establish posterior consistency. We place priors on the parameters that are fully flexible in the sense that the prior can generate densities arbitrarily close to any multivariate density (ruling out very irregular densities). We also show that the posterior will concentrate in arbitrarily small neighborhoods of the data generating density as a larger sample is collected. Two types of neighborhoods are considered: weak and total variation. Strong posterior consistency (total variation neighborhoods) is clearly a more interesting result. That said, we include details regarding weak posterior consistency as they are helpful in developing intuition regarding the regularity conditions used in the strong consistency results.

3.2.1 Weak Posterior Consistency

Consider a mixture density model f as in (3.3). Let $\mathcal{D}(\Re^m)$ denote the space of all densities on $\Re^m$. Let Π_f denote the prior induced on $\mathcal{D}(\Re^m)$ through the model and suitable priors on the parameters. Theorem 3.1 shows that Π_f satisfies the Kullback-Leibler (KL) condition at the true density f_t on $\Re^m$. That is, for any $\epsilon > 0$, $\Pi_f(K_\epsilon(f_t)) > 0$, where $K_\epsilon(f_t) = \{f \colon KL(f_t; f) < \epsilon\}$ denotes a ϵ-sized KL neighborhood of f_t and $KL(f_t; f) = \int \log \frac{f_t}{f} f_t dx$ is the KL divergence. As a result, using the Schwartz (1965) theorem, weak posterior consistency follows. That is, given a random sample $\mathbf{X}_n = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ i.i.d. f_t, the posterior

probability of any weak open neighborhood of f_t converges to 1 a.s. f_t.

Let $p(k)$ denote the prior distribution of k. We consider discrete priors that are supported on the set $\{0, \ldots, m\}$. Let $\pi_1(\boldsymbol{U}, \boldsymbol{\theta}|k)$ denote some joint prior distribution of $\boldsymbol{U}$ and $\boldsymbol{\theta}$ that has support on $\{(\boldsymbol{U}, \boldsymbol{\theta}) \in V_{k,m} \times \Re^m : \boldsymbol{U}'\boldsymbol{\theta} = \boldsymbol{0}\}$. As previously recommended, we consider a diagonal $\boldsymbol{\Sigma}_0 = diag(\sigma_1^2, \ldots, \sigma_k^2)$ and set a joint prior on the vector $\boldsymbol{\sigma} = (\sigma, \sigma_1, \ldots, \sigma_k) \in (\Re^+)^{k+1}$ that we denote with $\pi_2(\boldsymbol{\sigma}|k)$. Further, we assume that parameters ($\boldsymbol{U}$, $\boldsymbol{\theta}$), $\boldsymbol{\sigma}$, and P are jointly independent given k. That said, Theorem 3.1 can be easily adapted to other prior choices. We also consider the following regularity conditions on the true density f_t.

A1: $0 < f_t(\boldsymbol{x}) < A$ for some constant A for all $\boldsymbol{x} \in \Re^m$.

A2: $|\int \log\{f_t(\boldsymbol{x})\} f_t(\boldsymbol{x}) d\boldsymbol{x}| < \infty$.

A3: For some $\delta > 0$, $\int \log \frac{f_t(\boldsymbol{x})}{f_\delta(\boldsymbol{x})} f_t(\boldsymbol{x}) d\boldsymbol{x} < \infty$, where $f_\delta(\boldsymbol{x}) = \inf_{\boldsymbol{y}:\|\boldsymbol{y}-\boldsymbol{x}\|<\delta} f_t(\boldsymbol{y})$.

A4: For some $\alpha > 0$, $\int \|\boldsymbol{x}\|^{2(1+\alpha)m} f_t(\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Theorem 3.1. *Set the prior distributions for k, ($\boldsymbol{U}$, $\boldsymbol{\theta}$), $\boldsymbol{\sigma}$, and P to those described previously such that $p(m) > 0$, $\pi_2(\Re^+ \times (0, \epsilon)^m | k = m) > 0$ for any $\epsilon > 0$, and the conditional prior on P given $k = m$ contains P_{f_t} in its weak support. Then under assumptions* **A1**-**A4** *on f_t, the KL condition is satisfied by Π_f at f_t.*

Proof. The result follows if it can be proved that $\Pi_f(K_\epsilon(f_t)|k = m, \boldsymbol{U}) > 0$ for all $\epsilon > 0$ and $\boldsymbol{U} \in O(m)$, because then

$$\Pi_f(K_\epsilon(f_t)) \geq p(m) \int_{O(m)} \Pi_f(K_\epsilon(f_t)|k = m, \boldsymbol{U}) d\pi_1(\boldsymbol{U}|k = m) > 0$$

Now, given $k = m$ and $\boldsymbol{U}$, density (3.3) can be expressed as

$$f(\boldsymbol{x}; Q, \boldsymbol{\Sigma}) = \int_{\Re^m} N_m(\boldsymbol{x}; \boldsymbol{\nu}, \boldsymbol{\Sigma}) Q(d\boldsymbol{\nu}), \tag{3.5}$$

with $Q = P \circ \phi^{-1}$. Here $\phi(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{x}$, and $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Sigma}_0\boldsymbol{U}'$. The isomorphism $\phi : \Re^m \to \Re^m$ being continuous and surjective ensures the same for the mapping $P \mapsto Q$. This in turn ensures that under the assumptions of Theorem (3.1), the prior on P and $\boldsymbol{\sigma}$ induces a prior on Q that contains P_{f_t} in its weak support and an independent prior on $\boldsymbol{\Sigma}$ which induces a prior on its maximum eigen-value that contains 0 in its support. Then with a slight modification to the proof of Theorem 2 in Wu and Ghosal (2010), under assumptions **A1-A4** on f_t, we can show that f_t is in the KL support of Π_f. □

3.2.2 Strong Posterior Consistency

Using the density model (3.3) for f_t, Theorem 3.5 (see below) establishes strong posterior consistency. That is, the posterior probability of any total variation (or L_1 or strong) neighborhood of f_t converges to 1 almost surely or in probability, as the sample size tends to infinity. The priors on the parameters are chosen as in Section 3.2.1. To be more specific, the conditional prior on P given k ($k \geq 1$) is chosen to be a Dirichlet process $DP(w_k P_k)$ ($w_k > 0$, $P_k \in \mathcal{M}(\Re^k)$). The proof requires the following three Lemmas. The proof of Lemma (3.2) can be found in Barron (1988), while the proofs of Lemmas (3.3) and (3.4) are provided in the appendix.

In what follows $B_{r,m}$ refers to the set $\{\boldsymbol{x} \in \Re^m \colon \|\boldsymbol{x}\| \leq r\}$. For a subset $\mathcal{D}$ of densities and $\epsilon > 0$, the L_1-metric entropy $N(\epsilon, \mathcal{D})$ is defined as the logarithm of the minimum number of ϵ-sized (or smaller) L_1 subsets needed to cover $\mathcal{D}$.

Lemma 3.2. *Suppose that f_t is in the KL support of the prior Π_f on the density space $\mathcal{D}(\Re^m)$. For every $\epsilon > 0$, if we can partition $\mathcal{D}(\Re^m)$ as $\mathcal{D}_{n\epsilon} \cup \mathcal{D}_{n\epsilon}^c$ such that $N(\epsilon, \mathcal{D}_{n\epsilon})/n \longrightarrow 0$ and $Pr(D_{n\epsilon}^c | \mathbf{X}_n) \longrightarrow 0$ a.s. or in probability P_{f_t}, then the posterior probability of any L_1 neighborhood of f_t converges to 1 a.s. or in probability P_{f_t}.*

Lemma 3.3. *For positive sequences $h_n \to 0$ and $r_n \to \infty$ and $\epsilon > 0$, define a sequence of subsets of $\mathcal{D}(\Re^m)$ as*

$$\mathcal{D}_{n\epsilon} = \{f(\cdot; \boldsymbol{\Theta}) : \boldsymbol{\Theta} \in H_{n\epsilon}\},\ H_{n\epsilon} = \{\boldsymbol{\Theta} \colon \min(\boldsymbol{\sigma}) \geq h_n, \|\boldsymbol{\theta}\| \leq r_n, P(B_{r_n,k}^c) < \epsilon\}$$

with $f(\cdot; \boldsymbol{\Theta})$ as in (3.3). Set a prior on the density parameters as in Section 3.2.1. Assume that $supp(\pi_2(\cdot|k)) \subseteq [0, B]^{k+1}$ for some $B > 0$ for all $0 \leq k \leq m$. Then $N(\epsilon, \mathcal{D}_{n\epsilon}) \leq C(r_n/h_n)^m$ where C is a constant independent of n.

Lemma 3.4. *Set a prior as in Lemma 3.3 with a $DP(w_k P_k)$ prior on P given k, $k \geq 1$. Assume that the base probability P_k has a density p_k which is positive and continuous on $\Re^k$. Assume that there exist positive sequences $h_n \to 0$ and $r_n \to \infty$ such that*

$$\mathbf{B1} : \lim_{n\to\infty} n\delta_{kn}^{-1} h_n^{-k} \exp(-r_n^2/8B^2) = 0$$

holds where

$$\delta_{kn} = \inf\{p_k(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \Re^k,\ \|\boldsymbol{\mu}\| \leq B + r_n/2\},\ k = 1, \ldots, m.$$

Also assume that under the prior $\pi_2(\cdot|k)$ on $\boldsymbol{\sigma}$, $Pr(\min(\boldsymbol{\sigma}) < h_n|k)$ decays exponentially. Then under the assumptions of Theorem 3.1, for any $\epsilon > 0$, $k \geq 1$,

$$E_{f_t}\big\{Pr\big(P(B^c_{r_n,k}) \geq \epsilon \big| k, \mathbf{X}_n\big)\big\} \longrightarrow 0.$$

If **B1** *is strengthed to*

$$\mathbf{B1}' : \sum_{n=1}^{\infty} n\delta_{kn}^{-1} h_n^{-k} \exp(-r_n^2/8B^2) < \infty,$$

and the sequence r_n satisfies $\sum_{n=1}^{\infty} r_n^{-2(1+\alpha)m} < \infty$ with α as in Assumption **A4**, *then the conclusion can be strengthed to*

$$\sum_{n=1}^{\infty} E_{f_t}\big\{Pr\big(P(B^c_{r_n,k}) \geq \epsilon \big| k, \mathbf{X}_n\big)\big\} < \infty.$$

With these three Lemmas we are now able to state and proof the theorem that ensures strong posterior consistency is attained.

Theorem 3.5. *Consider a prior and sequences h_n and r_n for which the assumptions of Lemma 3.4 are satisfied. Further suppose that $n^{-1}(r_n/h_n)^m \longrightarrow 0$. Also assume that the sequence r_n and the prior $\pi_1(\cdot|k)$ on $(\boldsymbol{U}, \boldsymbol{\theta})$ satisfy the condition $Pr(\|\boldsymbol{\theta}\| > r_n|k)$ decays exponentially for $k \leq m-1$. Assume that the true density satisfies the conditions of Theorem 3.1. Then the posterior probability of any L_1 neighborhood of f_t converges to 1 in probability or almost surely depending on assumption* **B1** *or* **B1**$'$.

Proof. Theorem 3.1 implies that the KL condition is satisfied. Consider the partition $\mathcal{D}(\Re^m) = \mathcal{D}_{n\epsilon} \cup \mathcal{D}^c_{n\epsilon}$. Then $N(\epsilon, \mathcal{D}_{n\epsilon})/n \longrightarrow 0$. Write

$$Pr(\mathcal{D}^c_{n\epsilon}|\mathbf{X}_n) = Pr\big(\{f(.;\boldsymbol{\Theta}) : \boldsymbol{\Theta} \in H^c_{n\epsilon}\}\big|\mathbf{X}_n\big),$$

where

$$H^c_{n\epsilon} = \{\boldsymbol{\Theta} : \min(\boldsymbol{\sigma}) < h_n\} \cup \{\boldsymbol{\Theta} : \|\boldsymbol{\theta}\| > r_n\} \cup \{\boldsymbol{\Theta} : P(B^c_{r_n k}) > \epsilon\}.$$

The posterior probability of the first two sets above converge to 0 a.s. because the prior probability decays exponentially and the prior satisfies the KL condition. Note that

$$Pr\big(\{\boldsymbol{\Theta} : P(B^c_{r_n k}) > \epsilon\}\big|\mathbf{X}_n\big) \leq \sum_{j=1}^{m} Pr\big(\{\boldsymbol{\Theta} : P(B^c_{r_n k}) > \epsilon\}\big|\mathbf{X}_n, k = j\big)$$

and Lemma 3.4 implies that this probability converges to 0 in probability/a.s. based on Assumption **B1**/**B1**$'$.

Using Lemma 3.2, the result follows. □

Now we give an example of a prior that satisfies the conditions of Theorem 3.5. Any discrete distribution on $\{0, \ldots, m\}$ having m in its support can be used as the prior p for k. Given k ($k \geq 1$), we draw $\boldsymbol{U}$ from a density on $V_{k,m}$. Given k and $\boldsymbol{U}$, under π_1, $\boldsymbol{\theta}$ is drawn from a density on the vector-space $\mathcal{N}(\boldsymbol{U})$ if $k < m$. If $k = m$, then $\boldsymbol{\theta} = 0$. When $k < m$, we set $\boldsymbol{\theta} = r\tilde{\boldsymbol{\theta}}$ with r and $\tilde{\boldsymbol{\theta}}$ drawn independently from $\Re^+$ and the set $\{\tilde{\boldsymbol{\theta}} \in \Re^m : \|\tilde{\boldsymbol{\theta}}\| = 1, \tilde{\boldsymbol{\theta}}'\boldsymbol{U} = 0\}$ respectively. The scalar r^a is drawn from a Gamma density for appropriate $a > 0$. As a special case, a Gaussian density (conditioned to live on $\mathcal{N}(\boldsymbol{U})$) can be used for $\boldsymbol{\theta}$ when $\tilde{\boldsymbol{\theta}}$ is drawn uniformly, $a = 2$ and $r^2 \sim Gam(1, \sigma_0)$, $\sigma_0 > 0$. Then $\boldsymbol{\theta}$ has the density

$$\sigma_0^{-(m-k)} \exp \frac{-1}{2\sigma_0^2} \|\boldsymbol{\theta}\|^2 I(\boldsymbol{\theta}'\boldsymbol{U} = 0)$$

with respect to the volume form of $\mathcal{N}(\boldsymbol{U})$. Given k, $\boldsymbol{\sigma}$ follows π_2 which is supported on $[0, B]^{k+1}$ for appropriate $B > 0$. Under π_2, the coordinates of $\boldsymbol{\sigma}$ may be drawn independently with say, σ_j^{-2} following a Gamma density truncated to $[0, B]$. If reasonable, assuming $\sigma_1 = \ldots = \sigma_k = \sigma$ with σ^{-2} following a truncated Gamma density will simplify computations. That said, when $m \geq 2$ a Gamma distribution only satisfies the conditions of Theorem 3.1. To satisfy the conditions of Theorem 3.5 a truncated transformed Gamma density may be used. That is, for appropriate $b > 0$, we draw σ^{-b} from a Gamma density truncated to $[B^{-1}, \infty)$. Given k, $k \geq 1$, P follows a $DP(w_k P_k)$ prior. To get conjugacy, we may select P_k to be a Gaussian distribution on $\Re^k$ with covariance $\tau^2 \boldsymbol{I}_k$. With such a prior the conditions of Theorem 3.5 are satisfied if we choose a, b, τ and B such that $\tau^2 > 4B^2$, $a < 2(1+\alpha)m$ (with α as in **A4**) and $a^{-1} + b^{-1} < m^{-1}$. This result is available from Corollary 3.6 the proof of which is provided in the Appendix.

Corollary 3.6. *Assume that f_t satisfies Assumptions* **A1-A4**. *Let Π_f be a prior on the density space as in Theorem 3.5. Pick positive constants $a, b, \{\tau_k\}_{k=1}^m$ and B and set the prior as follows. Choose $\pi_1(.|k)$ such that for $k \leq m-1$, $\|\boldsymbol{\theta}\|^a$ follows a Gamma density. Pick $\pi_2(.|k)$ such that $\sigma, \sigma_1, \ldots, \sigma_k$ are independently and identicaly distributed with σ^{-b} following a Gamma density truncated to $[B^{-1}, \infty)$. Alternatively let $\sigma = \sigma_1 = \ldots = \sigma_k$ with σ distributed as above. For the $DP(w_k P_k)$ prior on P, $k \geq 1$, choose P_k to be a Normal density on $\Re^k$ with covariance $\tau_k^2 \boldsymbol{I}_k$. Then almost sure strong posterior consistency results if the constants satisfy $\tau_k^2 > 4B^2$, $a < 2(1+\alpha)m$ and $1/a + 1/b < 1/m$.*

A multivariate gamma prior on $\boldsymbol{\sigma}$ satisfies the requirements for weak but not strong posterior consistency (unless $m = 1$). However that does not prove that it is not eligible because Corollary 3.6 provides only sufficient conditions. Truncating the support of $\boldsymbol{\sigma}$ is not undesirable because for more precise fit we are

interested in low within cluster covariance which will result in sufficient number of clusters. However the transformation power b increases with m resulting in lower probability near zero which is undesirable when sample sizes are not high.

In Bhattacharya and Dunson (2011), a gamma prior is proved to to be eligible for a Gaussian mixture model (that is, $k = m$) as long as the hyperparameters are allowed to depend on sample size in a suitable way. However there it is assumed that f_t has a compact support. We expect the result to hold true in this context too.

3.3 Identifiability of Density Model Parameters

In many applications a principal modeling goal is actually estimating the subspace S and its dimension. In a classification or regression setting, this amounts to learning the principal subfeatures that explain the variability in the response. Before S can be uniquely estimated, it's identifiability must be established. That is, it must be shown that there exists a unique S corresponding to model (3.3). To this end, let P_f denote the distribution corresponding to f a mixture density as in (3.3). Then it follows that

$$P_f = N_m(\mathbf{0}, \mathbf{\Sigma}) * (P \circ \phi^{-1}), \tag{3.6}$$

with "$*$" denoting convolution. Now let $\Phi_P(t)$ be the characteristic function associated with P, then (3.6) implies that the characteristic function of f (or P_f) is

$$\Phi_f(t) = \exp(-1/2\boldsymbol{t}'\mathbf{\Sigma}\boldsymbol{t})\Phi_{P\circ\phi^{-1}}(t),\ \boldsymbol{t} \in \Re^m. \tag{3.7}$$

If a discrete P is employed, then (3.7) suggests that $\mathbf{\Sigma}$ and $P \circ \phi^{-1}$ can be uniquely determined from f. Recall that $\phi : \Re^k \longrightarrow \Re^m$ and $\phi(\Re^k) = S$. Further, if $Y \sim P$, then $P \circ \phi^{-1}$ is the distribution of $\phi(\boldsymbol{Y})$ with support on the k dimensional affine plane S.

In order to proceed we introduce the *affine support* of P denoted by asupp(P) and defined as the intersection of all affine subspaces of $\Re^k$ having probability 1. This actually turns out to be an affine subspace containing support(P) (but may be larger). To identify S and k we assume that asupp(P) is $\Re^k$. In other words, we use a prior for which P is discrete and asupp$(P) = \Re^k$ w.p. 1. An appropriate choice of prior for P given k would be the commonly used Dirichlet process with a full support base distribution. Then, from the nature of ϕ, asupp$(P \circ \phi^{-1})$ is an affine subspace of $\Re^m$ of dimension equal to that of asupp(P). Since asupp$(P \circ \phi^{-1})$ is identifiable, this implies that k is also identifiable as its dimension. Since

S contains asupp$(P \circ \phi^{-1})$ and has dimension equal to that of asupp$(P \circ \phi^{-1})$, $S = \text{asupp}(P \circ \phi^{-1})$. We have shown that the (sub) parameters $(\boldsymbol{\Sigma}, k, S, P \circ \phi^{-1})$ are identifiable once we set a full support discrete prior on P given k. Then $\boldsymbol{U}\boldsymbol{U}'$ and $\boldsymbol{\theta}$ are identifiable as the projection matrix and origin of S. However P and the coordinate choice ϕ (hence $\boldsymbol{U}$) are still non-identifiable. However, if we consider the structure $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Sigma}_0\boldsymbol{U}' + \sigma^2(\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}')$ with a diagonal $\boldsymbol{\Sigma}_0$ and impose some ordering on the diagonal entries of $\boldsymbol{\Sigma}_0$, then the columns of $\boldsymbol{U}$ become identifiable up to a change of signs as the eigen-rays.

3.4 Point Estimation for Subspace S

To obtain a Bayes estimate for the subspace S, one may choose an appropriate loss function and minimize the Bayes risk defined as the expectation of the loss over the posterior distribution. Any subspace is characterized by its projection matrix and origin. That is, the pair $(\boldsymbol{R}, \boldsymbol{\theta})$ where $\boldsymbol{R} \in M(m)$ and $\boldsymbol{\theta} \in \Re^m$ satisfy $\boldsymbol{R} = \boldsymbol{R}' = \boldsymbol{R}^2$ and $\boldsymbol{R}\boldsymbol{\theta} = \boldsymbol{0}$. We use $\mathcal{S}_m$ to denote the space of all such pairs. One particular loss function on $\mathcal{S}_m$ is

$$L_1((\boldsymbol{R}_1, \boldsymbol{\theta}_1), (\boldsymbol{R}_2, \boldsymbol{\theta}_2)) = \|\boldsymbol{R}_1 - \boldsymbol{R}_2\|^2 + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2,\ (\boldsymbol{R}_i, \boldsymbol{\theta}_i) \in \mathcal{S}_m,$$

where $\|\boldsymbol{A}\|^2 = \sum_{ij} a_{ij}^2 = \text{Tr}(\boldsymbol{A}\boldsymbol{A}')$ denotes the norm-squared of some matrix $\boldsymbol{A} = (a_{ij})$. Then a point estimate for $(\boldsymbol{R}, \boldsymbol{\theta})$ is the $(\boldsymbol{R}_1, \boldsymbol{\theta}_1)$ minimizing the posterior expectation of loss L_1 over $(\boldsymbol{R}_2, \boldsymbol{\theta}_2)$, provided there is a unique minimimizer.

If the goal is to estimate the directions of the subspace $(\boldsymbol{U})$, we may instead use the loss function

$$L_2((\boldsymbol{U}_1, w_1), (\boldsymbol{U}_2, w_2)) = \|\boldsymbol{U}_1 - \boldsymbol{U}_2\|^2 + (w_1 - w_2)^2,\ (\boldsymbol{U}_i, w_i) \in \mathcal{S}_{m2}.$$

Here the $m \times m$ matrix $\boldsymbol{U}_i$ has the first k columns as the directions of the corresponding subspace S_i, the $(k+1)$st gives the *direction* of the subspace origin $\boldsymbol{\theta}_i$ and the rest are set to the zero vector while $w_i = \|\boldsymbol{\theta}_i\|$. Therefore

$$\mathcal{S}_{m2} = \left\{ (\boldsymbol{U}, w) \in M(m) \times \Re^+ :\ \boldsymbol{U}'\boldsymbol{U} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \right\}.$$

Again the idea is to find the minimizer (if unique) $(\boldsymbol{U}_1, w_1)$ of the expected value of L_2 under the posterior distribution of $(\boldsymbol{U}_2, w_2)$ and set the estimated subspace dimension k as the rank of $\boldsymbol{U}_1$ minus 1, the principal directions consisting of the first k columns of $\boldsymbol{U}_1$ and the origin as w_1 times the last column. Since the k

orthonormal directions of the subspace are only identifiable as rays, one may even look at the loss

$$L_3((\boldsymbol{U},\boldsymbol{\theta}_1),(\boldsymbol{V},\boldsymbol{\theta}_2)) = \sum_{j=1}^{m} \|\boldsymbol{U}_j\boldsymbol{U}_j' - \boldsymbol{V}_j\boldsymbol{V}_j'\|^2 + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2,$$

where

$$(\boldsymbol{U},\boldsymbol{\theta}_1),(\boldsymbol{V},\boldsymbol{\theta}_2) \in \mathcal{S}_{m3} = \left\{ (\boldsymbol{U},\boldsymbol{\theta}) \in M(m)\times\Re^m :\ \boldsymbol{U}'\boldsymbol{U} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},\ \boldsymbol{U}'\boldsymbol{\theta} = \boldsymbol{0} \right\}.$$

Theorems 3.7 and 3.8 (proofs of which can be found in the appendix) derive expressions for minimimizers of the risk functions corresponding to L_1 and L_2 and present conditions for their uniqueness. In what follows we use P_n to denote the posterior distribution of the parameters given the sample which is assumed to have finite second order moments. For a matrix A, by $A_{(k)}$ we shall denote the submatrix of A consisting of its first k columns.

Theorem 3.7. *Let* $f_1(\boldsymbol{R},\boldsymbol{\theta}) = \int_{(\boldsymbol{R}_2,\boldsymbol{\theta}_2)} L_1((\boldsymbol{R},\boldsymbol{\theta}),(\boldsymbol{R}_2,\boldsymbol{\theta}_2))dP_n(\boldsymbol{R}_2,\boldsymbol{\theta}_2)$, $(\boldsymbol{R},\boldsymbol{\theta}) \in \mathcal{S}$. *This function is minimized by* $\boldsymbol{R} = \sum_{j=1}^{k} \boldsymbol{U}_j\boldsymbol{U}_j'$ *and* $\boldsymbol{\theta} = (\boldsymbol{I}-\boldsymbol{R})\bar{\boldsymbol{\theta}}$ *where* $\bar{\boldsymbol{R}} = \int_{M(m)} \boldsymbol{R}_2 dP_n(\boldsymbol{R}_2)$ *and* $\bar{\boldsymbol{\theta}} = \int_{\Re^m} \boldsymbol{\theta}_2 dP_n(\boldsymbol{\theta}_2)$ *are the posterior means of* $\boldsymbol{R}_2$ *and* $\boldsymbol{\theta}_2$ *respectively,* $2\bar{\boldsymbol{R}} - \bar{\boldsymbol{\theta}}\bar{\boldsymbol{\theta}}' = \sum_{j=1}^{m} \lambda_j \boldsymbol{U}_j\boldsymbol{U}_j'$, $\lambda_1 \geq \ldots \geq \lambda_m$ *is a s.v.d. of* $2\bar{\boldsymbol{R}} - \bar{\boldsymbol{\theta}}\bar{\boldsymbol{\theta}}'$, *and* k *minimizes* $k - \sum_{j=1}^{k} \lambda_j$ *on* $\{0,\ldots,m\}$. *The minimizer is unique if and only if there is a unique* k *minimizing* $k - \sum_{j=1}^{k} \lambda_j$ *and* $\lambda_k > \lambda_{k+1}$ *for that* k.

Theorem 3.8. *Let* $f_2(\boldsymbol{U},w) = \int_{(\boldsymbol{U}_2,w_2)} L_2((\boldsymbol{U},w),(\boldsymbol{U}_2,w_2))dP_n(\boldsymbol{U}_2,w_2)$, $(\boldsymbol{U},w) \in \mathcal{S}_{m2}$. *Let* $\bar{w}$ *and* $\bar{\boldsymbol{U}}$ *denote the posterior means of* w_2 *and* $\boldsymbol{U}_2$ *respectively. Then* f_2 *is minimized by* $w = \bar{w}$ *and any* $\boldsymbol{U} = [\boldsymbol{U}_1, 0]$, *where* $\boldsymbol{U}_1 \in V_{k+1,m}$ *satisfies* $\bar{\boldsymbol{U}}_{(k+1)} = \boldsymbol{U}_1(\bar{\boldsymbol{U}}'_{(k+1)}\bar{\boldsymbol{U}}_{(k+1)})^{1/2}$, *and* k *minimizes* $g(k) = k - 2\mathrm{Tr}(\bar{\boldsymbol{U}}'_{(k+1)}\bar{\boldsymbol{U}}_{(k+1)})^{1/2}$ *over* $\{0,\ldots,m-1\}$. *The minimizer is unique if and only if there is a unique* k *minimizing* g *and* $\bar{\boldsymbol{U}}_{(k+1)}$ *has full rank for that* k.

3.5 Nonparametric Classification with Feature Coordinate Selection

We know turn our attention to specifying a kernel for Y and ultimately the joint of $(Y,\boldsymbol{X})$. Because the association between $\boldsymbol{X}$ and Y may not be causal, it is natural to model $\boldsymbol{X}$ and Y jointly and employ the conditional that is induced through the joint to model $Y|\boldsymbol{X}$. Beyond being conceptually pleasing, this strategy provides a coherent method of dealing with missing observations and is quite flexible in the types of data that can be accommodated. In classification, Y is categorical and takes on values from $\{1,\ldots,c\}$.

Because of this, a multinomial kernel (denoted by $M_c(y;\boldsymbol{\nu}) = \prod_{\ell=1}^{c} \nu_\ell^{I[y=\ell]}$) would be a natural choice for Y. Using the specified kernels for $\boldsymbol{X}$ and Y consider the following model

$$(Y, \boldsymbol{X}) \sim f(y, \boldsymbol{x}) = \int_{\Re^k \times S_c} N_m(\boldsymbol{x}; \phi(\boldsymbol{\mu}), \boldsymbol{\Sigma}) M_c(y; \boldsymbol{\nu}) P(d\boldsymbol{\mu}\, d\boldsymbol{\nu}), \tag{3.8}$$

with $S_c = \{\boldsymbol{\nu} \in [0,1]^c \colon \sum \nu_\ell = 1\}$ denoting the $c-1$ dimensional simplex. Note that (3.8) is a generalization of (3.3) and (3.4) along the lines of the joint model proposed in Dunson and Bhattacharya (2011), though they focus on kernels for predictors on models that accommodate non-Euclidean manifolds and there is no dimensionality reduction.

Since $\boldsymbol{X}$ is high dimensional it is possible (and fairly common) that the estimation of $Y|\boldsymbol{X}$ through $(Y, \boldsymbol{X})$ is dominated by the marginal on $\boldsymbol{X}$. To avoid this we desire to identify a few "important" coordinates of $\boldsymbol{X}$ and model $(Y, \boldsymbol{X})$ only through those coordinates. This would induce a conditional that depends on only a few coordinates of $\boldsymbol{X}$ (thus simultaneously performing dimension reduction and feature coordinate selection). The remaining coordinates of $\boldsymbol{X}$ can be modeled independently as equal variance Gaussians (though preliminary studies indicate that the subspace estimation and prediction are robust to a true joint distribution having 'non-signal' predictors that are not predictive of Y).

to this end consider an isometric transformation on $\boldsymbol{X}$. (Note that an isometric transformation can be used with out loss of generality and it provides some benefit regarding coordinate inversion.) That is, one can locate a $k \leq m$ and $\boldsymbol{U} \in V_{k,m}$ such that

$$(Y, \boldsymbol{U}'\boldsymbol{X}) \sim f_1(y, \boldsymbol{U}'\boldsymbol{x}) = \int_{\Re^k \times S_c} N_k(\boldsymbol{U}'\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_0) M_c(y; \boldsymbol{\nu}) P(d\boldsymbol{\mu}\, d\boldsymbol{\nu}),\ \boldsymbol{U}'\boldsymbol{x} \in \Re^k. \tag{3.9}$$

Additionally for some $\boldsymbol{\theta} \in \Re^m$ and $\boldsymbol{V} \in V_{m-k,m}$ satisfying $\boldsymbol{V}'\boldsymbol{U} = \boldsymbol{0}$ and $\boldsymbol{\theta}'\boldsymbol{U} = \boldsymbol{0}$, the orthogonal residual is modeled as

$$\boldsymbol{V}'\boldsymbol{X} \sim N_{m-k}(\boldsymbol{V}'\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_{m-k}) \tag{3.10}$$

independently of $(Y, \boldsymbol{U}'\boldsymbol{X})$. With such a structure, the joint distribution of $(Y, \boldsymbol{X})$ becomes (3.8) where

$$\phi : \Re^k \to \Re^m,\ \phi(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{x} + \boldsymbol{\theta},\ \boldsymbol{U} \in V_{k,m},\ \boldsymbol{\theta} \in \Re^m,\ \boldsymbol{U}'\boldsymbol{\theta} = \boldsymbol{0},$$
$$\boldsymbol{\Sigma} = \boldsymbol{U}(\boldsymbol{\Sigma}_0 - \sigma^2 \boldsymbol{I}_k)\boldsymbol{U}' + \sigma^2 \boldsymbol{I}_m,\ \boldsymbol{\Sigma}_0 \in M^+(k), \sigma^2 \in \Re^+.$$

The conditional density $Y = y|\boldsymbol{X} = \boldsymbol{x}$ can be expressed as

$$p(y|x;\boldsymbol{\Theta}) = \frac{\int_{\Re^k \times S_c} N_k(\boldsymbol{U}'\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}_0) M_c(y;\boldsymbol{\nu}) P(d\boldsymbol{\mu}\, d\boldsymbol{\nu})}{\int_{\Re^k \times S_c} N_k(\boldsymbol{U}'\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}_0) P(d\boldsymbol{\mu}\, d\boldsymbol{\nu})} \tag{3.11}$$

with parameters $\boldsymbol{\Theta} = (k, \boldsymbol{U}, \boldsymbol{\Sigma}_0, P, \boldsymbol{\theta}, \sigma^2)$. A draw from the posterior of $\boldsymbol{\Theta}$ given model (3.8) will provide a draw from (3.11).

When P is discrete (which is a standard choice), $Y|\boldsymbol{X}, \boldsymbol{\Theta}$ can be thought of as a weighted c dimensional multinomial probability vector with the weights depending on $\boldsymbol{X}$ only through the selected k-dimensional coordinates $\boldsymbol{U}'\boldsymbol{X}$. For example, if $P = \sum_{j=1}^{\infty} w_j \delta_{(\boldsymbol{\mu}_j, \boldsymbol{\nu}_j)}$, then

$$p(y|\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{j=1}^{\infty} \tilde{w}_j(\boldsymbol{U}'\boldsymbol{x}) M_c(y;\boldsymbol{\nu}_j) \tag{3.12}$$

where $\tilde{w}_j(\boldsymbol{U}'\boldsymbol{x}) = \frac{w_j N_k(\boldsymbol{U}'\boldsymbol{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_0)}{\sum_{i=1}^{\infty} w_i N_k(\boldsymbol{U}'\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_0)}$ and $\boldsymbol{U}'\boldsymbol{x} \in \Re^k$ for $j = 1, \ldots, \infty$. We refer to (3.12) as the principal subspace classifier (PSC).

The above is easily adapted to a regression setting by considering a low dimensional response $\boldsymbol{Y} \in \Re^l$ and replacing the multinomial kernel used for Y with a Gaussian kernel. In this setting the joint model becomes

$$(\boldsymbol{Y}, \boldsymbol{X}) \sim \int_{\Re^k \times \Re^l} N_m(\boldsymbol{x}; \phi(\boldsymbol{\mu}), \boldsymbol{\Sigma}_x) N_l(\boldsymbol{y}; \boldsymbol{\psi}, \boldsymbol{\Sigma}_y) P(d\boldsymbol{\mu}\, d\boldsymbol{\psi}), \tag{3.13}$$

which produces the following conditional model

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Theta}) = \frac{\int_{\Re^k \times \Re^l} N_k(\boldsymbol{U}'\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}_0) N_l(\boldsymbol{y};\boldsymbol{\psi},\boldsymbol{\Sigma}_y) P(d\boldsymbol{\mu}\ d\boldsymbol{\psi})}{\int_{\Re^k \times \Re^l} N_k(\boldsymbol{U}'\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}_0) P(d\boldsymbol{\mu}\, d\boldsymbol{\psi})}. \tag{3.14}$$

For a discrete P this conditional distribution becomes

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{j=1}^{\infty} \tilde{w}_j(\boldsymbol{U}'\boldsymbol{x}) N_l(\boldsymbol{y};\boldsymbol{\psi}_j,\boldsymbol{\Sigma}_y). \tag{3.15}$$

which is a mixture whose weights depend on $\boldsymbol{X}$ only through its k-dimensional coordinates $\boldsymbol{U}'\boldsymbol{X}$. As the regression model is a straightforward modification of the classifier, we focus on the classification case for sake of brevity.

4 Posterior Computation

Since sampling independently from the distribution of $\boldsymbol{\Theta} = (k, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\Sigma}_0, \sigma, P)$ conditioned on the *iid* realizations $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ is not feasible, we resort to obtaining MCMC draws. As will be seen the computation required for fitting the model is fairly straight forward as a Gibbs sampler is all that is required.

As a prior for P, we use a Dirichlet process (DP) (i.e., $P \sim DP(w_0 P_0)$) with $P_0 = N_k(\boldsymbol{m}_\mu, \boldsymbol{S}_\mu)$ and $w_0 = 1$. We employ the Sethuraman (1994)'s stick breaking representation of the Dirichlet process so that $P = \sum_{j=1}^{\infty} w_j \delta_{\boldsymbol{\mu}_j}$ where $\boldsymbol{\mu}_j$ is drawn *iid* from P_0 and $w_j = v_j \prod_{\ell<j}(1 - v_\ell)$ with $v_j \sim Beta(1, w_0)$. After introducing cluster labels $S_1, \ldots, S_n$, the likelihood becomes

$$f(\boldsymbol{x}; \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\Sigma}_0, \sigma, \boldsymbol{\mu}, \boldsymbol{S}, \boldsymbol{w}, \nu) = \prod_{i=1}^{n} w_{S_i} N_m(\boldsymbol{x}_i; \boldsymbol{U}\boldsymbol{\mu}_{S_i} + \boldsymbol{\theta}, \boldsymbol{\Sigma}) M_c(y_i; \boldsymbol{\nu}_{S_i})$$

Assuming *a priori* independence between the elements of $\boldsymbol{\Theta}$ we select commonly used conjugacy preserving prior distributions for parameters $\boldsymbol{\Sigma}_0$ and σ along with latent variables $\boldsymbol{\mu}$, $\boldsymbol{S}$, $\boldsymbol{w}$ and ν . A von Mises-Fisher prior distribution is used for $(\boldsymbol{U}|\boldsymbol{\theta})$ while a truncated normal is used for $\boldsymbol{\theta}$. For sake of brevity, we only highlight the particularly novel parts of the MCMC algorithm (updating $\boldsymbol{U}$ and $\boldsymbol{\theta}$). Details regarding full conditionals of the of the remaining parameters and latent variables are provided in the Appendix.

Updating $\boldsymbol{U}$:

Let $\pi(\boldsymbol{U}) \propto \text{etr}\{\boldsymbol{A}'\boldsymbol{U}\}$ denote a von Mises-Fisher prior distribution for $\boldsymbol{U} \in V_{k,m}$. It can be shown that the full conditional of $\boldsymbol{U}$ is

$$\begin{aligned}
[\boldsymbol{U}|-] &\propto \exp\{tr\big[1/2(\sigma^{-2}\boldsymbol{I}_k - \boldsymbol{\Sigma}_0^{-1})\boldsymbol{U}'(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i')\boldsymbol{U} + \boldsymbol{\Sigma}_0^{-1}(\sum_{i=1}^{n}\boldsymbol{\mu}_{S_i}\boldsymbol{x}_i')\boldsymbol{U}\big]\}\pi(\boldsymbol{U})I[\boldsymbol{U}'\boldsymbol{\theta} = \boldsymbol{0}] \\
&\propto \text{etr}\{\boldsymbol{F}_1'\boldsymbol{U} + \boldsymbol{F}_2\boldsymbol{U}'\boldsymbol{F}_3\boldsymbol{U}\}\text{etr}\{\boldsymbol{A}'\boldsymbol{U}\}I[\boldsymbol{U}'\boldsymbol{\theta} = \boldsymbol{0}], \\
&\propto \text{etr}\{(\boldsymbol{F}_1 + \boldsymbol{A})'\boldsymbol{U} + \boldsymbol{F}_2\boldsymbol{U}'\boldsymbol{F}_3\boldsymbol{U}\}I[\boldsymbol{U}'\boldsymbol{\theta} = \boldsymbol{0}] \qquad (4.1)
\end{aligned}$$

where $\boldsymbol{F}_1 = (\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{\mu}_{S_i}')\boldsymbol{\Sigma}_0^{-1}$, $\boldsymbol{F}_2 = \frac{1}{2}(\sigma^{-2}\boldsymbol{I}_k - \boldsymbol{\Sigma}_0^{-1})$, and $\boldsymbol{F}_3 = \sum_{i=1}^{n}(\boldsymbol{x}_i\boldsymbol{x}_i')$. Thus the full conditional of $\boldsymbol{U}$ is a Bingham-von Mises-Fisher distribution constrained to lie in $\mathcal{N}(\boldsymbol{\theta}')$. Strategies for sampling from an unconstrained Bingham-von Mises-Fisher are developed in Hoff (2009) and can be extended for use here. Using a change of variable technique, it can be shown that for any orthonormal basis $\boldsymbol{N}$ of $\mathcal{N}(\boldsymbol{\theta}')$ the

distribution of $\tilde{\boldsymbol{U}} = \boldsymbol{N}'\boldsymbol{U}$ is

$$[\tilde{\boldsymbol{U}}|-] \propto \text{etr}\{[\boldsymbol{N}'(\boldsymbol{F}_1 + \boldsymbol{A})]'\tilde{\boldsymbol{U}} + \boldsymbol{C}\tilde{\boldsymbol{U}}'\boldsymbol{N}'\boldsymbol{B}\boldsymbol{N}\tilde{\boldsymbol{U}}\} I[\tilde{\boldsymbol{U}}'\tilde{\boldsymbol{U}} = \boldsymbol{I}_k], \tag{4.2}$$

which is an unconstrained Bingham-von Mises-Fisher distribution. Therefore, a draw from the full conditional of $\boldsymbol{U}$ is obtained by carrying out the following steps

1. Obtain $\boldsymbol{N}$ an othonormal basis of $\mathcal{N}(\boldsymbol{\theta}')$.
2. Using ideas from Hoff (2009) to sample $\tilde{\boldsymbol{U}}$ from (4.2).
3. Set $\boldsymbol{U} = N\tilde{\boldsymbol{U}}$.

Updating $\boldsymbol{\theta}$:

A conjugate prior for $\boldsymbol{\theta}$ conditioned on $\boldsymbol{U}$ is $N(\boldsymbol{m}_\theta, \boldsymbol{S}_\theta) I[\boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{U}')]$. The full conditional of $\boldsymbol{\theta}$ with this prior is $[\boldsymbol{\theta}|-] \sim N(\boldsymbol{m}^*_\theta, \boldsymbol{S}^*_\theta) I[\boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{U}')]$ where $\boldsymbol{S}^*_\theta = (n\boldsymbol{\Sigma}^{-1} + \boldsymbol{S}_\theta^{-1})^{-1}$ and $\boldsymbol{m}^*_\theta = \boldsymbol{S}^*_\theta(n\boldsymbol{\Sigma}^{-1}\bar{x} + \boldsymbol{S}_\theta^{-1}\boldsymbol{m}_\theta)$. One can sample from this truncated normal distribution by employing the same change of variable technique used to sample from $[\boldsymbol{U}|-]$. Specifically, let $\boldsymbol{M}$ denote an orthonormal basis of $\mathcal{N}(\boldsymbol{U}')$. Then $[\tilde{\boldsymbol{\theta}} = \boldsymbol{M}'\boldsymbol{\theta}|-]$ (the coordinates associated with the projection $\boldsymbol{M}\boldsymbol{M}'\boldsymbol{\theta}$) follows an unconstrained multivariate normal distribution. Therefore the following steps can be followed to sample from $\boldsymbol{\theta}$

1. Obtain $\boldsymbol{M}$ an orthonormal basis for $\mathcal{N}(\boldsymbol{U}')$
2. Sample $\tilde{\boldsymbol{\theta}}$ from a $m-k$ dimensional Normal distribution with mean $(\boldsymbol{M}'\boldsymbol{S}^{*-1}_\theta\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{S}^{*-1}_\theta\boldsymbol{m}^*_\theta$ and covariance $(\boldsymbol{M}'\boldsymbol{S}^{*-1}_\theta\boldsymbol{M})^{-1}$.
3. Set $\boldsymbol{\theta} = \boldsymbol{M}\tilde{\boldsymbol{\theta}}$

Though the algorithm converges quite quickly due to the orthogonality of the projection coordinates, reasonable starting values can decrease the number of MCMC iterates discarded as burn-in and therefore may be desirable. For $\boldsymbol{U}$, the first k eigen-vectors of the sample covariance matrix can be used. For $\boldsymbol{\theta}$ one may use $(\boldsymbol{I}_m - \boldsymbol{U}_s\boldsymbol{U}_s')\bar{\boldsymbol{x}}$ where $\boldsymbol{U}_s$ denotes the starting value for $\boldsymbol{U}$. The initial labels (S_i) and coordinate cluster means $(\boldsymbol{\mu}_j)$ can be obtained by applying a k-means algorithm to $\boldsymbol{U}_s'\boldsymbol{x}_i$.

5 Simulations and Examples

We provide a few simulated and real data examples to demonstrate density estimation, classification, subspace estimation and highlight parameter interpretations available from the PSD and PSC.

5.1 Density Estimation and Prediction Simulation Study

To assess density estimation and prediction performance we employ two data generating mechanisms. First we generate m-dimensional $\boldsymbol{X}$ vectors using a finite mixture $\boldsymbol{X} \sim \sum_{h=1}^{c+1} \pi_h N_m(\boldsymbol{\eta}_h, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{\eta}_h$ is a vector of zeros save for the hth entry which is 1 and σ^2 is the bandwidth. Y is generated using (3.12) along with an orthogonal basis of the matrix that projects $\boldsymbol{X}$ onto the c-dimensional plane. This method of generating data will be referred to as the mixture data generator (MDG). Secondly we use a factor model $\boldsymbol{X} \sim N_m(\boldsymbol{0}, \boldsymbol{\Omega})$ where $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is a diagonal matrix whose inverse diagonal values are generated with $Gam(1, 0.25)$ (a Gamma distribution with mean 4) and the entries of $\boldsymbol{\Lambda}$ are chosen as described in Bhattacharya and Dunson (2010) making $\boldsymbol{\Omega}$ sparse. Y is created by first generating a $(m+1)$-dimensional $\boldsymbol{X}$ vector and setting $Y = I[\boldsymbol{X}_{[1]} > 0]$ ($\boldsymbol{X}_{[1]}$ denotes the first element of $\boldsymbol{X}$). We refer to this data generating scenario as factor model data generator (FDG). In addition to multiple data generating schemes we also change the feature space dimension ($m = 50, 100$) and subspace dimension ($k = 2, 5$) but fix the sample size at $n = 100$ and bandwidth at $\sigma^2 = 0.1$ (for MDG). For each factor level combination 25 data set replicates were created. The PSD and PSC models were fit to these replicate data sets using 1,000 MCMC iterates after discarding the first 10,000 as burn-in and thinning by 50.

Table 1: Results of the Kullback-Leibler type distance comparing estimated densities from each of the procedures considered in the simulation study to the density used to generate data.

Data generating mechanism	k used in generating data	m	PSD $k=2$	PSD $k=5$	Fin Mix	Inf Mix
Finite Mixture	2	50	151.51	153.72	367.51	287.37
		100	287.56	289.51	477.47	541.43
	5	50	294.03	298.39	484.70	567.70
		100	525.16	531.21	690.92	712.64
Factor Model	2	50	204.71	220.68	288.91	317.56
		100	247.57	255.96	285.38	286.17
	5	50	5312.87	5296.59	5479.63	5324.01
		100	9498.37	9466.23	10075.46	9725.70

As competitors to PSD we consider a finite mixture $f(\boldsymbol{x}) = \sum_{h=1}^{c} \pi_h N_m(\boldsymbol{x}; \boldsymbol{\mu}_h, \sigma^2 \boldsymbol{I}_m)$ and an infinite mixture $f(\boldsymbol{x}) = \sum_{h=1}^{\infty} \pi_h N_m(\boldsymbol{x}; \boldsymbol{\mu}_h, \sigma^2 \boldsymbol{I}_m)$. These were chosen because of their flexibility and accuracy in estimating densities. To compare the density estimates from the mixtures to those produced by the PSD

the following Kullback-Leibler type distance is used

$$\frac{1}{D}\sum_{d=1}^{D}\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{\ell=1}^{100}\log f_0(\boldsymbol{x}^*_{\ell d})-\sum_{\ell=1}^{100}\log \hat{f}_t(\boldsymbol{x}^*_{\ell d})\right). \tag{5.1}$$

Here f_0 denotes the true density function, d is an index for the $D = 25$ replicate data sets, $\boldsymbol{x}^*_{\ell d}$ is the ℓth out of sample observation generated from the dth data set and $\hat{f}_t$ is the estimated density. Values for (5.1) can be found in Table 1 for the various data generating scenarios. The values in columns "PSD $k = 2$" and "PSD $k = 5$" are the results of (5.1) using the PSD model with k fixed at 2 and 5 respectively. Results from the finite mixture and infinite mixture are under the columns "Fin Mix" and "Inf Mix".

It appears that PSD does a better job of estimating the true density relative to the mixtures. This is true regardless of the data generating mechanism and the true subspace dimension. This is somewhat expected for PSD models that "over-fit" the data in the sense that the subspace dimension used in the model is larger than the true subspace dimension. However, this was quite unexpected in the case of "under-fitting" (value of k used in the model is smaller than true subspace dimension).

To compare the classification performance of PSC we consider k nearest neighbor (KNN), mixture discriminant analysis (MDA), and support vector machine (SVM). KNN and SVM are very accurate algorithmic based classifiers while MDA is a well known model based classifier. The `R` R Development Core Team (2010) functions `knn` Venables and Ripley (2002), `mda` Hastie and Tibshirani (2009), and `svm` Dimitriadou et al. (2011) were employed to implement KNN, MDA, and SVM. For KNN the neighborhood size corresponding to the best out of sample prediction was used. Similarly, for MDA the number of components that produced the best out of sample prediction was used. The four methodologies were compared using out of sample prediction error rates. To investigate the influence that "over-fitting" and "under-fitting" might have on PSC predictions, the PSC model was fit using $k = 2, 5, 10$. Results can be found in Table 2. Values under the columns "PSC $k = \cdot$" correspond to the out of sample error prediction rate under the PSC model with fixed $k = \cdot$. The values under KNN, MDA, and SVM represent the out of sample prediction error rates for the respective procedures.

It seems fairly evident from the simulation study that the PSC is an able classifier. When generating data using the mixture the PSC attains the lowest out of sample prediction error regardless of wether the true subspace dimension is used in fitting the model or not. However, as the subspace dimension increased it appears that the consequences of under-fitting are more precarious than over-fitting (a similar phenomena in most models developed for prediction).

Table 2: Results of the out of sample prediction error rates.

Data generating mechanism	k used in generating data	m	PSC $k=2$	PSC $k=5$	PSC $k=10$	KNN	MDA	SVM
Finite Mixture	2	50	0.028	0.029	0.029	0.166	0.060	0.066
		100	0.032	0.032	0.033	0.250	0.147	0.128
	5	50	0.151	0.060	0.058	0.238	0.123	0.198
		100	0.189	0.073	0.068	0.323	0.220	0.289
Factor Model	2	50	0.210	0.210	0.210	0.240	0.300	0.220
		100	0.190	0.180	0.180	0.220	0.470	0.220
	5	50	0.210	0.160	0.190	0.330	0.220	0.170
		100	0.237	0.198	0.201	0.372	0.286	0.199

5.2 Illustrations Using Real Datasets

As an empirical example, consider the so called Brain Computer Interface (BCI) data set from the third BCI competition. This data set consists of a single person performing 400 trials. In each one movements with the left hand or the right hand were imagined and the EEG was recorded from 39 electrodes. An autoregressive model of order 3 was Þt to each of the resulting 39 time series. A trial is then represented by the total of $117 = 39 \times 3$ dimensional feature space. The goal is to classify each trial as left or right hand movements using the 117 features. After standardizing the data, we randomly selected 200 observations to serve as testing data. To select a dimension k we considered out of sample prediction error and area under the ROC curve. Since low out of sample prediction error and large area under the curve are desirable the $k \in 1, \ldots, 25$ that maximized their difference was used (which turned out to be $k = 3$). The resulting out of sample prediction error rate for PSC was 0.205 compared to 0.51 for KNN, 0.25 for MDA and 0.23 for SVM. Therefore, in this empirical example, PSC is an adequate classifier in a moderately large feature space.

Though we have shown that the method provides comparable prediction (relative to commonly used alternatives), the PSC model has the advantage of providing interpretable parameters. To investigate this further, consider the Wisconsin Breast Cancer data set which is available in the `mlbench` (Leisch and Dimitriadou 2010) `R` package. In this data set the response is breast cancer diagnosis while the covariates are nine nominal variables describing some type of breast tissue cell characteristic. The values of the nine covariates were determined by medical experts. To fit the PSC model, $k = 3$ was used (which was arrived at using the same procedure described for the BCI data set). With this k, the PSC produced an out of sample error rate of 0.017 which is smaller than the error rate for KNN (0.035), MDA (0.028) and SVM (0.028). More importantly, particular interest in this study centered on determining the influence that each tumor attribute

had on classification. The nine attributes (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis) are all related to a lump being benign or not. Using results from Theorem 3.7 the estimated principal directions are found in Table 3. (Recall these directions are unique only up to a sign under certain criteria outlined in Section 4.5.)

Table 3: The $k = 3$ principal directions of the Breast Cancer data set along with the row norms

Variable	$U_{[,1]}$	$U_{[,2]}$	$U_{[,3]}$	norm
clump thickness	-0.294	0.233	0.453	0.588
uniformity of cell size	-0.399	-0.132	-0.189	0.460
uniformity of cell shape	-0.395	-0.102	0.0172	0.408
marginal adhesion	-0.314	-0.007	-0.477	0.571
single epithelial cell size	-0.231	-0.181	-0.307	0.424
bare nuclei	-0.450	0.713	0.101	0.849
bland chromatin	-0.295	-0.032	-0.194	0.354
normal nucleoli	-0.376	-0.587	0.543	0.883
mitosis	-0.121	-0.173	-0.305	0.371

The principal directions can be interpreted similarly to how principal components are interpreted (see the end of Section 3.1). From this perspective, the first principal direction is a function of all nine covariates and their weights are all similar. The next principal direction excludes marginal adhesion and bland chromatin while grouping clump thickness and bare nuclei with more weight placed on the later. Another method that can be used to assess the relative importance of each variable and also provide a means of grouping the variables is to calculate the norm associated with each row of $\boldsymbol{U}$ (this criteria is invariant to choice of $\boldsymbol{U}$). These values can be found under the "norm" column of Table 3. It appears that bare nuclei and normal nucleoli form a group, clump thickness and marginal adhesion form another, and uniformity of cell size, uniformity of cell shape and single epithelial cell size form a third.

5.3 Collaborative Perinatal Project Application

We now revisit the Collaborative Perinatal Project (CPP) sub-study that was briefly described in the introduction. A principal aim of the study was to investigate the influence that exposure to certain chemicals might have on preterm and small-for-gestational-age babies at birth. To investigate possible associations, samples from 2380 subjects in the CPP were taken and assayed for various compounds (e.g., the insecticides DDT and Dieldrin, metabolite DDE, and albumin). Other subject specific characteristics were also recorded

(e.g., race, length of gestation). In addition to identifying exposures that are important in explaining the responses of interest, there is substantial interest in being able to characterize the joint effect of a combination or mixture of the exposures. These could be used to create an index of an individual's exposure burden which could be used to determine a mothers risk to adverse gestational outcomes. The PSC seems particularly well suited to carry out the study objectives. For sake of simplicity we only considered 40 variables that were identified to be of particular interest. Among them are lipid adjusted exposure measurements to the metabolite DDE, the insecticides DDT, Dieldrin, Mirex, and Oxychlorodane and many measurements related to poychlorinated biphenyls (PCBs). Also, variables such as race, age, and length of gestation were included. The response was an indicator of pre-term birth.

We randomly partitioned the data set into training (2000 individuals) and testing (380 individuals) data sets. Because the adverse response (preterm birth) is rare, it is not useful to use out of sample prediction error rate as a model assessment tool or a criterion to choose a dimension (regular-term birth was always predicted which is the case for SVM as well). Therefore, only out of sample area under the ROC curve was used to select k which resulted in $k = 3$. To approximate the posterior distribution of $\boldsymbol{\Theta}$ given observations, 1000 MCMC iterates were collected after discarding the first 100,000 as burn-in and thinning by 100. Convergence was assessed graphically using two independently run MCMC chains. Once the 1000 MCMC iterates were collected $\boldsymbol{U}$ was estimated using the ideas in Section 3.4. Upon looking at these principal directions a little closer a few interesting groups of predictors appeared. Table 4 provides results in tabular form and Figure 2 provides a graphical representation.

The first principal direction contains 20 relevant predictors with all the 9 lipid unadjusted PCB metabolites being grouped together. (e.g., they had similar magnitudes.) The second principal direction identified race as the variable with the largest magnitude and DDE (both lipid adjusted and not) as a group. The third principal direction contains 15 predictors with the sum of the total lipid-adjustd PCB's having the largest magnitude. Other predictors that were fairly clearly grouped were lipid adjusted PCB 138 and 175, and total cholesterol. Considering the row norms, PCB 153 and 138 adjusted for lipid amounts are clearly most relevant while Mirex both adjusted for lipids and not are clearly the least.

6 Conclusions

This article has proposed a novel methodology for nonparametric Bayesian learning of an affine subspace underlying potentially high-dimensional data. Clearly, there is a need for flexible methods for dimension-

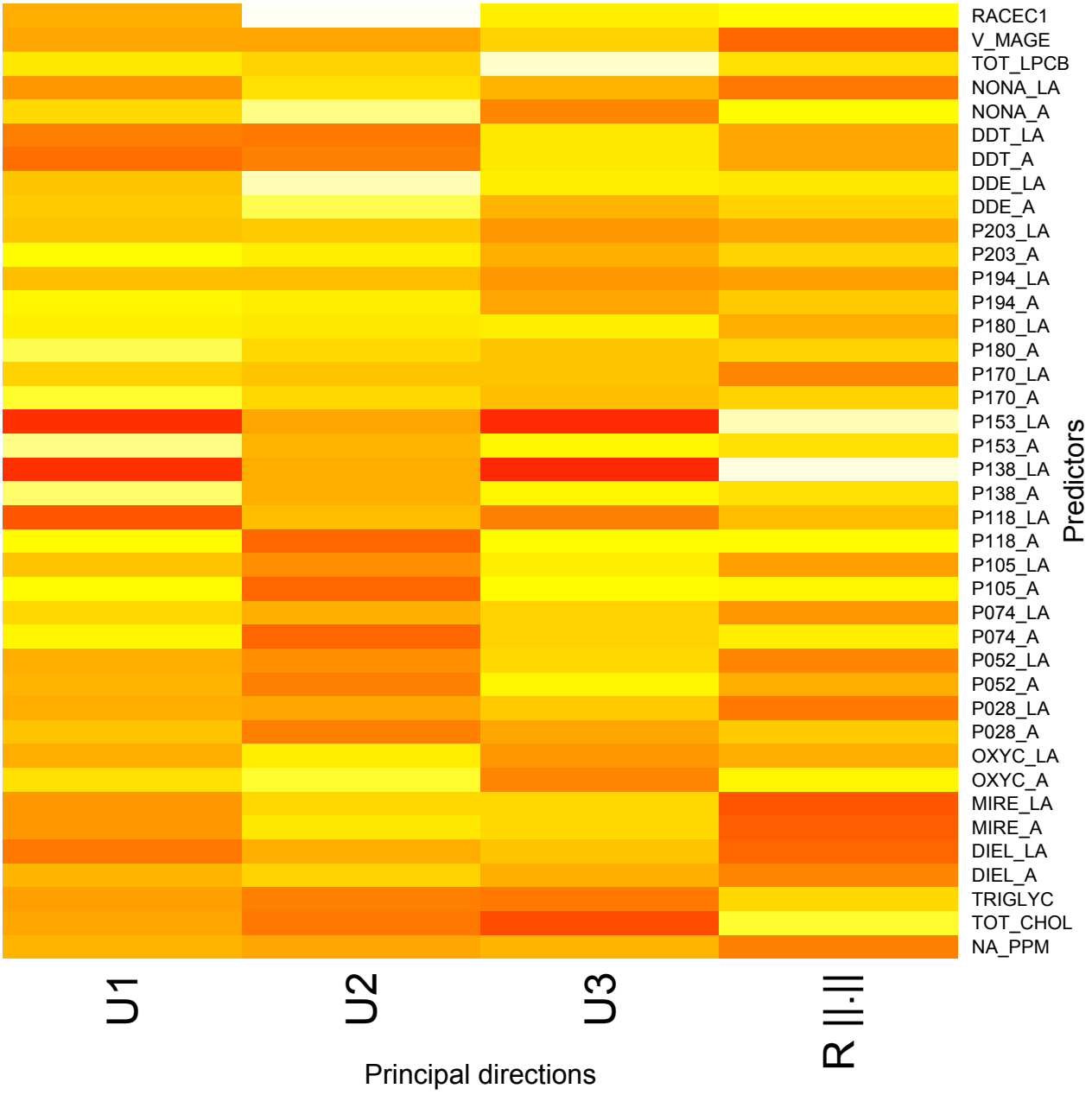

Figure 2: Heat map of the 3 principal directions of the Perinatal data set along with the row norms

Table 4: The 3 principal directions of the Collaborative Perinatal Project sub-study along with the row norms

Variable	$U_{[,1]}$	$U_{[,2]}$	$U_{[,3]}$	norm
NA_PPM	0.081	-0.083	-0.093	0.149
TOT_CHOL	0.053	-0.211	-0.332	0.397
TRIGLYC	0.040	-0.184	-0.241	0.306
DIEL_A	0.092	0.044	-0.119	0.157
DIEL_LA	-0.025	-0.068	-0.072	0.103
MIRE_A	0.021	0.087	-0.009	0.090
MIRE_LA	0.036	0.063	-0.016	0.074
OXYC_A	0.181	0.210	-0.215	0.351
OXYC_LA	0.075	0.123	-0.166	0.219
PCB028_A	0.125	-0.199	-0.138	0.272
PCB028_LA	0.077	-0.096	-0.038	0.129
PCB052_A	0.089	-0.190	0.067	0.220
PCB052_LA	0.072	-0.146	-0.018	0.164
PCB074_A	0.228	-0.258	-0.024	0.345
PCB074_LA	0.167	-0.067	-0.026	0.182
PCB105_A	0.238	-0.246	0.090	0.354
PCB105_LA	0.120	-0.145	0.046	0.193
PCB118_A	0.239	-0.258	0.099	0.366
PCB118_LA	-0.096	-0.023	-0.230	0.250
PCB138_A	0.295	-0.077	0.069	0.313
PCB138_LA	-0.168	-0.059	-0.455	0.489
PCB153_A	0.306	-0.053	0.066	0.318
PCB153_LA	-0.161	-0.083	-0.423	0.460
PCB170_A	0.271	0.050	-0.091	0.291
PCB170_LA	0.143	-0.003	-0.064	0.156
PCB180_A	0.285	0.047	-0.058	0.295
PCB180_LA	0.206	0.088	0.045	0.228
PCB194_A	0.217	0.116	-0.137	0.282
PCB194_LA	0.096	-0.032	-0.171	0.199
PCB203_A	0.236	0.121	-0.127	0.294
PCB203_LA	0.119	0.022	-0.172	0.211
DDE_A	0.139	0.230	-0.094	0.285
DDE_LA	0.114	0.299	0.047	0.323
DDT_A	-0.049	-0.199	0.034	0.207
DDT_LA	-0.011	-0.212	0.021	0.213
NONA_A	0.159	0.261	-0.210	0.371
NONA_LA	0.026	0.084	-0.095	0.129
TOT_LPCB	0.199	0.040	0.234	0.310
V_MAGE	0.059	-0.082	-0.034	0.106
RACEC1	0.070	0.375	0.038	0.383

ality reduction that avoid parametric assumptions. In this context, the Bayesian paradigm has substantial advantages over commonly used machine learning, computer science and frequentist statistical methods that obtain a point estimate of the subspace or manifold which the data are concentrated near. As there is unavoidably substantial uncertainty in subspace or manifold learning, it is important to fully account for this uncertainty to avoid misleading inferences and obtain appropriate measures of uncertainty in estimating densities, performing predictions and identifying important predictors. We accomplish this in a Bayesian manner by placing a probability model over the space of affine subspaces, while developing a simple and efficient computational algorithm relying on Gibbs sampling to estimate the subspace. The model is theoretically proved to be highly flexible and posterior consistency is achieved under appropriate prior choices. The proposed model and computational algorithm should be broadly useful beyond the density estimation and classification settings we have considered. In addition to building efficient classifiers, the methodology provides insight regarding predictors (or mixtures of them) that are influential in explaining the variability in the response, information that applied scientists often consider valuable.

Acknowlegements: This work was partially supported by Award Number R01ES017436 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

Appendices

A Proofs

As a reminder in what follows $B_{r,m}$ refers to the set $\{\boldsymbol{x} \in \Re^m \colon \|\boldsymbol{x}\| \leq r\}$. For a subset $\mathcal{D}$ of densities and $\epsilon > 0$, the L_1-metric entropy $N(\epsilon, \mathcal{D})$ is defined as the logarithm of the minimum number of ϵ-sized (or smaller) L_1 subsets needed to cover $\mathcal{D}$.

A.1 Proof of Lemma (3.3)

Proof. Any density f in $\mathcal{D}_{n\epsilon}$ can be expressed as $\int_{\Re^m} N_m(\boldsymbol{\nu}, \boldsymbol{\Sigma}) Q(d\boldsymbol{\nu})$ with $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Sigma}_0\boldsymbol{U}' + \sigma^2(\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}')$, $Q = P \circ \phi^{-1}$, $\phi(\boldsymbol{x}) = \boldsymbol{U}\boldsymbol{x}$, and $(k, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \sigma, P) \in H_{n\epsilon}$. The assumption on π_2 and $H_{n\epsilon}$ will imply that $\boldsymbol{\Sigma}$ has all its eigenvalues in $[h_n^2, A^2]$.

We also claim that $Q(B^c_{\sqrt{2}r_n,m}) < \epsilon$. To see this, note that $\|\phi(\boldsymbol{\mu})\|^2 = \|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\theta}\|^2 \leq 2r_n^2$ whenever $\|\boldsymbol{\mu}\| \leq r_n$ and $\|\boldsymbol{\theta}\| \leq r_n$. Hence $B_{r_n,k} \subseteq \phi^{-1}(B_{\sqrt{2}r_n,m})$ if $\|\boldsymbol{\theta}\| \leq r_n$. Therefore $\epsilon > P(B^c_{r_n,k}) \geq P\big((\phi^{-1}(B_{\sqrt{2}r_n,m}))^c\big) = P \circ \phi^{-1}\big(B^c_{\sqrt{2}r_n,m}\big)$ for all $(P,\boldsymbol{\theta}) \in H_{n\epsilon}$. Hence the claim follows. Therefore

$$\mathcal{D}_{n\epsilon} \subseteq \tilde{\mathcal{D}}_{n\epsilon} = \{f = \int N_m(\boldsymbol{\nu},\boldsymbol{\Sigma})Q(d\boldsymbol{\nu}) : Q(B^c_{\sqrt{2}r_n,m}) < \epsilon,\ \lambda(\boldsymbol{\Sigma}) \in [h_n^2, A^2]\},$$

$\lambda(\boldsymbol{\Sigma})$ denoting the eigenvalues of $\boldsymbol{\Sigma}$. From Lemma 1 of Wu and Ghosal (2010), it follows that $N(\epsilon, \tilde{\mathcal{D}}_{n\epsilon}) \leq C(r_n/h_n)^m$ and this completes the proof. □

A.2 Proof of Lemma (3.4)

The proof is similar in scope to the proof of Lemma 2 in Wu and Ghosal (2010). Throughout the proof, C will denote constant independent of n.

Proof. Given $k, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}$ and $\boldsymbol{\mu}_n = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n)$ iid P with $\boldsymbol{X}_i \sim N_m\big(\phi(\boldsymbol{\mu}_i), \boldsymbol{\Sigma}\big)$, $i = 1, \ldots, n$, mutually independent and independent of P. Hence

$$Pr\big(P(B^c_{r_n,k}) \geq \epsilon \big| k, \mathbf{X}_n\big) = E\big(Pr\big(P(B^c_{r_n,k}) \geq \epsilon \big| k, \boldsymbol{\mu}_n\big)\big| k, \mathbf{X}_n\big).$$

From Ferguson (1973), given $\boldsymbol{\mu}_n$ and k, for $A \subseteq \Re^k$, $P(A) \sim Beta\big(w_k P_k(A) + N(A), w_k(1-P_k) + n - N(A)\big)$ where $N(A) = \sum_{i=1}^n I_{\{\boldsymbol{\mu}_i \in A\}}$. Hence using the Markov inequality,

$$Pr\big(P(B^c_{r_n,k}) \geq \epsilon \big| k, \boldsymbol{\mu}_n\big) \leq \frac{w_k P_k(B^c_{r_n,k}) + N(B^c_{r_n,k})}{\epsilon(n + w_k)}.$$

Therefore

$$E\big(Pr\big(P(B^c_{r_n,k}) \geq \epsilon \big| k, \mathbf{X}_n\big) \leq \frac{w_k P_k(B^c_{r_n,k})}{\epsilon(n + w_k)} + \frac{1}{\epsilon(n + w_k)} \sum_{i=1}^n Pr\big(\boldsymbol{\mu}_i \in B^c_{r_n,k} \big| k, \mathbf{X}_n\big).$$

Denote the above two terms as T_1 and T_2. Then $E_{f_t} T_1 = T_1 \longrightarrow 0$ as $r_n \to \infty$. Under the marginal prior given k, $\boldsymbol{\mu}_n$ has an exchangable distribution $\pi_n(\boldsymbol{\mu}_n|k)$ on $(\Re^k)^n$ (see Ferguson (1973)). Also since $\mathbf{X}_n$ are iid given f_t, it follows that

$$E_{f_t}(T_2) = \frac{n}{\epsilon(n + w_k)} E_{f_t}\big\{Pr\big(\boldsymbol{\mu}_1 \in B^c_{r_n,k} \big| k, \mathbf{X}_n\big)\big\}.$$

Now

$$Pr\big(\boldsymbol{\mu}_1 \in B^c_{r_n,k}\big|k,\mathbf{X}_n\big) \le Pr\big(\boldsymbol{\mu}_1 \in B^c_{r_n,k}, \min(\boldsymbol{\sigma}) > h_n\big|k,\mathbf{X}_n\big) + Pr(\min(\boldsymbol{\sigma}) \le h_n\big|k,\mathbf{X}_n).$$

The last term above converges to 0 a.s. by the assumption on π_2. Hence to complete the proof, it remains to show that

$$E_{f_t}\big\{Pr\big(\boldsymbol{\mu}_1 \in B^c_{r_n,k}, \min(\boldsymbol{\sigma}) > h_n\big|k,\mathbf{X}_n\big)\big\} \longrightarrow 0 \text{ as } n \to \infty.$$

To compute the probability in above, we denote by $\pi_{1n}(\boldsymbol{\mu}_1|\boldsymbol{\mu}_{-1},k)$ the conditional distribution of $\boldsymbol{\mu}_1$ given $\boldsymbol{\mu}_{-1} = (\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_n)$, and by $\pi_{-1n}(\boldsymbol{\mu}_{-1}|k)$ the marginal distribution of $\boldsymbol{\mu}_{-1}$ under the joint π_n. Then

$$Pr\big(\boldsymbol{\mu}_1 \in B^c_{r_n,k}, \min(\boldsymbol{\sigma}) > h_n\big|k,\mathbf{X}_n\big) = A(\mathbf{X}_n)/B(\mathbf{X}_n)$$

where $A(\mathbf{X}_n) =$

$$\int\limits_{\min(\boldsymbol{\sigma})>h_n, \|\boldsymbol{\mu}_1\|>r_n} \prod_{i=1}^{n} N_m(\boldsymbol{X}_i;\phi(\boldsymbol{\mu}),\boldsymbol{\Sigma}) d\pi_{1n}(\boldsymbol{\mu}_1|\boldsymbol{\mu}_{-1},k) d\pi_{-1n}(\boldsymbol{\mu}_{-1}|k) d\pi_1(\boldsymbol{U},\boldsymbol{\theta}|k) d\pi_2(\boldsymbol{\sigma}|k)$$

and $B(\mathbf{X}_n) =$

$$\int \prod_{i=1}^{n} N_m(\boldsymbol{X}_i;\phi(\boldsymbol{\mu}),\boldsymbol{\Sigma}) d\pi_{1n}(\boldsymbol{\mu}_1|\boldsymbol{\mu}_{-1},k) d\pi_{-1n}(\boldsymbol{\mu}_{-1}|k) d\pi_1(\boldsymbol{U},\boldsymbol{\theta}|k) d\pi_2(\boldsymbol{\sigma}|k).$$

We use $E_{f_t}\{A(\mathbf{X}_n)/B(\mathbf{X}_n)\} \le$

$$\sup_{\boldsymbol{X}_1 \in B_{r_n/2,m}} \frac{A(\mathbf{X}_n)}{B(\mathbf{X}_n)} \int_{B_{r_n/2,m}} f_t(\boldsymbol{x}) d\boldsymbol{x} + \int_{B^c_{r_n/2,m}} f_t(\boldsymbol{x}) d\boldsymbol{x}. \tag{A.1}$$

and upper bound the terms in above.

First we upper bound $A(\mathbf{X}_n)$ when $\|\boldsymbol{X}_1\| \le r_n/2$. We express $N_m(\boldsymbol{X}_1;\phi(\boldsymbol{\mu}_1),\boldsymbol{\Sigma})$ as

$$N_k(\boldsymbol{U}'\boldsymbol{X}_1;\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_0)$$

and note that $\|\boldsymbol{X}_1\| \leq r_n/2$, $\|\boldsymbol{\mu}_1\| > r_n$ and $h_n < \sigma_j \leq A\ \forall j \leq k$ implies

$$N_k(\boldsymbol{U}'\boldsymbol{X}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0) \leq C h_n^{-k} \exp \frac{-r_n^2}{8A^2}.$$

Therefore $A(\mathbf{X}_n) \leq$

$$\begin{aligned} C h_n^{-k} \exp \frac{-r_n^2}{8A^2} \int (\sigma^{-2})^{\frac{m-k}{2}} \exp \frac{-1}{2\sigma^2} (\boldsymbol{X}_1 - \boldsymbol{\theta})'(\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}')(\boldsymbol{X}_1 - \boldsymbol{\theta}) \\ \prod_{i=2}^{n} N_m(\boldsymbol{X}_i; \phi(\boldsymbol{\mu}_i), \boldsymbol{\Sigma}) d\pi_{-1n}(\boldsymbol{\mu}_{-1}|k) d\pi_1(\boldsymbol{U}, \boldsymbol{\theta}|k) d\pi_2(\boldsymbol{\sigma}|k). \end{aligned} \tag{A.2}$$

Next we lower bound $B(\mathbf{X}_n)$ when $\boldsymbol{X}_1 \in B_{r_n/2,m}$. The conditional distribution π_{1n} can be expressed as $\frac{1}{w_k+n-1} \sum_{i=2}^{n} \delta_{\boldsymbol{\mu}_i} + \frac{w_k}{w_k+n-1} P_k$ (see Ferguson (1973)). Hence $B(\mathbf{X}_n) \geq$

$$\frac{w_k}{w_k + n - 1} \int \prod_{i=1}^{n} N_m(\boldsymbol{X}_i; \phi(\boldsymbol{\mu}_i), \boldsymbol{\Sigma}) p_k(\boldsymbol{\mu}_1) d\boldsymbol{\mu}_1 d\pi_{-1n}(\boldsymbol{\mu}_{-1}|k) d\pi_1(\boldsymbol{U}, \boldsymbol{\theta}|k) d\pi_2(\boldsymbol{\sigma}|k).$$

Now

$$\int N_k(\boldsymbol{U}'\boldsymbol{X}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0) p_k(\boldsymbol{\mu}_1) d\boldsymbol{\mu}_1 \geq \int_S N_k(\boldsymbol{U}'\boldsymbol{X}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0) p_k(\boldsymbol{\mu}_1) d\boldsymbol{\mu}_1$$

where

$$S = \{\boldsymbol{\mu}_1 : \sum_{l=1}^{k} \sigma_l^2 (\boldsymbol{U}_k'\boldsymbol{X}_1 - \boldsymbol{\mu}_1)_l^2 \leq 1\}.$$

For $\boldsymbol{\mu}_1 \in S$, $N_k\big(\boldsymbol{U}'\boldsymbol{X}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0\big) \geq \prod_1^k \sigma_j^{-1} e^{-1/2}$ and $p_k(\boldsymbol{\mu}_1) \geq \delta_{kn}$ with δ_{kn} defined in the Lemma. Therefore

$$\int_S N_k(\boldsymbol{U}'\boldsymbol{X}_1; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0) p_k(\boldsymbol{\mu}_1) d\boldsymbol{\mu}_1 \geq C\delta_{kn} \prod_1^k \sigma_j^{-1} \int_S d\boldsymbol{\mu}_1 = C\delta_{kn}$$

and hence when $\|\boldsymbol{X}_1\| \leq r_n/2$, $B(\mathbf{X}_n) \geq$

$$\begin{aligned} C n^{-1} \delta_{kn} \int (\sigma^{-2})^{\frac{m-k}{2}} \exp \frac{-1}{2\sigma^2} (\boldsymbol{X}_1 - \boldsymbol{\theta})'(\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}')(\boldsymbol{X}_1 - \boldsymbol{\theta}) \prod_{i=2}^{n} N_m(\boldsymbol{X}_i; \phi(\boldsymbol{\mu}_i), \boldsymbol{\Sigma}) \\ d\pi_{-1n}(\boldsymbol{\mu}_{-1}|k) d\pi_1(\boldsymbol{U}, \boldsymbol{\theta}|k) d\pi_2(\boldsymbol{\sigma}|k). \end{aligned} \tag{A.3}$$

Combining (A.2) and (A.3), we get

$$\sup_{\|\boldsymbol{X}_1\| \leq r_n/2} \frac{A(\mathbf{X}_n)}{B(\mathbf{X}_n)} \leq C n \delta_{kn}^{-1} h_n^{-k} \exp(-r_n^2/8A^2).$$

Plug this in (A.1) to conclude $E_{f_t}\{A(\mathbf{X}_n)/B(\mathbf{X}_n)\} \leq$

$$Cn\delta_{kn}^{-1}h_n^{-k}\exp(-r_n^2/8A^2) + Pr_{f_t}(\|\boldsymbol{X}\| > r_n/2) \tag{A.4}$$

which converges to zero by assumption.

Under assumption **B1'** and $\sum r_n^{-2(1+\alpha)m} < \infty$ the sequence in (A.4) has a finite sum which results in the stronger conclusion. This completes the proof. □

A.3 Proof of Corollary (3.6)

Proof. By Theorem 3.5, to show a.s. strong posterior consistency, we need to get positive sequences r_n and h_n which satisfy

$$n^{-1}(r_n/h_n)^m \longrightarrow 0,\ \sum r_n^{-2(1+\alpha)m} < \infty,\ \text{and} \tag{A.5}$$

$$\sum_{n=1}^{\infty} n\delta_{kn}^{-1}h_n^{-k}\exp(-r_n^2/8A^2) < \infty, \tag{A.6}$$

and the prior probabilities $Pr(\|\boldsymbol{\theta}\| > r_n|k)$ and $Pr(\min(\boldsymbol{\sigma}) < h_n|k)$ decay exponentially. Set $r_n = n^{1/a}$ and $h_n = n^{-1/b}$. Then (A.5) is clearly satisfied. By the choice of p_k, $k \geq 1$, it is easy to check that $\delta_{kn} \geq C\exp\frac{-r_n^2}{2\tau_k^2}$ with C denoting positive constants independent of n all throughout. Then (A.6) is clearly satisfied because of the assumption $\tau_k^2 > 4A^2$. Also because $\|\boldsymbol{\theta}\|^a$ follows a Gamma distribution given k, $k \leq m-1$, the probability $Pr(\|\boldsymbol{\theta}\| > r_n|k)$ can be upper bounded by $C\exp(-\lambda r_n^a)$ for some $\lambda > 0$. This decays exponentially with $r_n = n^{1/a}$. Lastly, it remains to check that $Pr(\min(\boldsymbol{\sigma}) < h_n|k)$, decays exponentially. When the coordinates of $\boldsymbol{\sigma}$ are all equal, the probability can be upper bounded by $C\exp(-\lambda h_n^{-b})$ for some $\lambda > 0$. This decays exponentially with $h_n = n^{-1/b}$. In case the coordinates are iid, the probability can be upper bounded by $Cn\exp(-\lambda h_n^{-b})$ which also decays exponentially by the choice of h_n. □

A.4 Proof of Theorem (3.7)

Proof. Simplify f_1 as

$$\begin{aligned} f_1(\boldsymbol{R},\boldsymbol{\theta}) &= f_1(\bar{\boldsymbol{R}},\bar{\boldsymbol{\theta}}) + \|\boldsymbol{R}-\bar{\boldsymbol{R}}\|^2 + \|\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\|^2 \\ &= f_1(\bar{\boldsymbol{R}},\bar{\boldsymbol{\theta}}) + \|\boldsymbol{R}-\bar{\boldsymbol{R}}\|^2 + \|\boldsymbol{R}\bar{\boldsymbol{\theta}}\|^2 + \|(\boldsymbol{I}-\boldsymbol{R})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})\|^2 \\ &\geq f_1(\bar{\boldsymbol{R}},\bar{\boldsymbol{\theta}}) + \|\boldsymbol{R}-\bar{\boldsymbol{R}}\|^2 + \|\boldsymbol{R}\bar{\boldsymbol{\theta}}\|^2. \end{aligned} \tag{A.7}$$

Equality holds in (A.7) iff $\boldsymbol{\theta} = (\boldsymbol{I} - \boldsymbol{R})\bar{\boldsymbol{\theta}}$. Then

$$f_1(\boldsymbol{R}, \boldsymbol{\theta}) = k - \mathrm{Tr}\{(2\bar{\boldsymbol{R}} - \bar{\boldsymbol{\theta}}\bar{\boldsymbol{\theta}}')\boldsymbol{R}\} + C$$

where k =Rank($\boldsymbol{R}$) and C denotes something not depending on $\boldsymbol{R}, \boldsymbol{\theta}$. From the proof of Proposition 11.1 Bhattacharya and Bhattacharya (2011, In Press), given k one can show that the value of $\boldsymbol{R}$ minimizing f_1 above is $\sum_{j=1}^{k} \boldsymbol{U}_j \boldsymbol{U}_j'$ and the minimizer is unique iff $\lambda_k > \lambda_{k+1}$. Then

$$f_1(\boldsymbol{R}, \boldsymbol{\theta}) = k - \sum_{j=1}^{k} \lambda_j + C.$$

Now one needs to find the k minimizing the above risk which is as mentioned. This completes the proof. □

A.5 Proof of Theorem (3.8)

Proof. The minimizer $w = \bar{w}$ is obvious. Then

$$f_2(\boldsymbol{U}, \bar{w}) = \|\boldsymbol{U} - \bar{\boldsymbol{U}}\|^2 + C = k_1 - 2\mathrm{Tr}\bar{\boldsymbol{U}}_{(k_1)}'\boldsymbol{U}_{(k_1)} + C,$$

k_1 being the rank of $\boldsymbol{U}$ and C symbolizing any constant not depending on $\boldsymbol{U}$. For k_1 fixed, it is proved in Theorem 10.2 Bhattacharya and Bhattacharya (2011, In Press) that the minimizer $\boldsymbol{U}$ is as in the theorem. It is unique iff $\bar{\boldsymbol{U}}_{(k_1)}'\bar{\boldsymbol{U}}_{(k_1)}$ is invertible. Plug that $\boldsymbol{U}$ and the risk function becomes, as a function of k_1,

$$f_3(k_1) = k_1 - 2\mathrm{Tr}(\bar{\boldsymbol{U}}_{(k_1)}'\bar{\boldsymbol{U}}_{(k_1)})^{1/2}.$$

We find the value of k_1 between 1 and m minimizing f_3 and set $k = k_1 - 1$. □

B Full Conditionals and MCMC algorithm details

We provide the remainder of the full conditionals. After updating $\boldsymbol{U}$ and $\boldsymbol{\theta}$ as described in Section 5, a complete Gibbs sampler can be constructed by cycling through the following 6 steps on an individual basis.
Step 1: Update S_i for $i = 1, 2, \ldots, n$ by sampling from the following conditional posterior distribution

$$Pr(S_i = j|-) \propto w_j \exp\left\{-1/2(\boldsymbol{\mu}_j'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_j - 2\boldsymbol{\mu}_j'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{U}'\boldsymbol{x}_i)\right\} \prod_{\ell=1}^{c} \nu_{j\ell}^{I[y_i=\ell]}$$

for $j = 1, \ldots, \infty$. To make the total number of states finite the block Gibbs sampler of Ishwaran and James (2001) may be implemented. Alternatively, the slice sampling ideas described in Yau et al. (2011), Walker (2007), or Kalli et al. (2011) could be used. The remainder of the algorithm is described from the perspective of using a block Gibbs sampler which requires truncating the number of atoms to N.

Step 2: Update the DP atom weights by setting $w_j = v_j \prod_{l=1}^{j-1}(1 - v_l)$, $j = 1, \ldots, N$ after drawing

$$[v_l|-] \sim Beta(1 + n_j, w_0 + \sum_i I(S_i > j))$$

with $n_j = \sum_i I(S_i = j)$ and setting $v_N = 1$.

Step 3: Update the DP atoms $\{(\boldsymbol{\mu}_j, \boldsymbol{\nu}_j) : j = 1, \ldots, N\}$ independently by sampling from

$$[\boldsymbol{\mu}_j|-] \sim N_k(\boldsymbol{m}^*_\mu, \boldsymbol{S}^*_\mu),$$

where $\boldsymbol{S}^*_\mu = (n_j \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{S}_\mu^{-1})^{-1}$ and $\boldsymbol{m}^*_\mu = S^*_\mu(\boldsymbol{U}'\boldsymbol{\Sigma}_0^{-1} \sum_{i:S_i=j} \boldsymbol{x}_i + \boldsymbol{S}_\mu^{-1}\boldsymbol{m}_\mu)$. Update the $\boldsymbol{\nu}_j$'s by sampling from

$$[\boldsymbol{\nu}_j|-] \sim Dir(a^*_1, \ldots, a^*_c),$$

where $a^*_\ell = \sum_{i=1}^n I[y_i = \ell, S_i = j] + a_\ell$ for $\ell = 1, \ldots, c$.

Step 4: Using a $\sigma^{-2} \sim \text{Ga}(a, b)$ prior, σ^{-2} can be updated using

$$[\sigma^{-2}|-] \sim \text{Ga}(\frac{1}{2}n(m-k) + a, b + \frac{1}{2}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i + \frac{n}{2}\boldsymbol{\theta}'\boldsymbol{\theta} - \frac{1}{2}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{U}\boldsymbol{U}'\boldsymbol{x}_i - \boldsymbol{\theta}'\sum_{i=1}^n \boldsymbol{x}_i)$$

Under the simplifying assumption that $\boldsymbol{\Sigma}_0 = \sigma^2 I_k$ the full conditional of σ^{-2} becomes

$$[\sigma^{-2}|-] \sim \text{Ga}(\frac{1}{2}nm + a, b + \frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{\mu}_{S_i} - \boldsymbol{\theta})'(\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{\mu}_{S_i} - \boldsymbol{\theta}))$$

Step 5: Using a truncated Gamma distribution for σ_j^{-2} (i.e., $\sigma_j^{-2} \sim Gam(a, b)I[\sigma_j^{-2} \in [0, A]]$) allows one to update σ_j^{-2} using the following truncated Gamma distribution.

$$[\sigma_j^{-2}|-] \sim \text{GAM}(\frac{n}{2} + a, b + \frac{1}{2}\sum_{i=1}^n (\boldsymbol{U}'\boldsymbol{x}_i - \boldsymbol{\mu}_{S_i})_j^2) I[\sigma_j^{-2} \in [0, A]].$$

References

Barron, A. R. (1988), "The Exponential Convergence of posterior probabilities with implications for Bayes estimators of density functions," *Technical Report*, 7.

Bhattacharya, A. and Bhattacharya, R. (2011, In Press), *Nonparametric Statistics on Manifolds with Applications to Shape Spaces, IMS Monograph Series*, Cambridge University Press.

Bhattacharya, A. and Dunson, D. B. (2010), "Sparse Bayesian infinite factor models," *Biometrika*.

— (2011), "Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds," *Annals of the Institute of Statistical Mathematics*, 1–28.

— (2012), "Nonparametric Bayes classification and hypothesis testing on manifolds," *Journal of Multivariate Analysis*, 111, 1–19.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., and Carin, L. (2010), "Compressive Sensing on Manifolds Using a Nonparametric Mixture of Factor Analyzers: Algorithm and Performance Bounds," *IEEE Signal Processing*, 58, 6140–6155.

Chikuse, Y. (2003), *Statistics on Special Manifolds, Lecture Notes in Statistics*, vol. 174, New York: Springer-Verlag.

Cook, R. D. and Weisberg, S. (1991), "Sliced Inverse Regression for Dimension Reduction: Comment," *Journal of the American Statistical Association*, 86, 328–332.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks," *Machine Learning*, 20, 273–297.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., , and Weingessel, A. (2011), *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, r package version 1.6.

Dunson, D. B. and Bhattacharya, A. (2011), "Nonparametric Bayes regression and classification through mixtures of product kernels," in *Bayesian Statistics*, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., vol. 9, pp. 145–164.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *Annals of Statististics*, 1, 209–230.

Hastie, T. and Tibshirani, R. (2009), *mda: Mixture and flexible discriminant analysis*, r package version 0.4-1.

Hastie, T., Tibshirani, R., and Friedman, J. (2008), *The Elements of Statistical Learning*, Springer, 2nd ed.

Hoff, P. D. (2007), "Model Averaging and Dimension Selection for the Singular Value Decomposition," *Journal of the American Statistical Association*, 102, 674–685.

— (2009), "Simulation of the Matrix Bingham-von Mises-Fisher Distribution, With Applications to Multivariate Relational Data," *Journal of Computational and Graphical Statistics*, 18, 438–356.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–73.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

Leisch, F. and Dimitriadou, E. (2010), *mlbench:: Machine Learning Benchmark Problems.*, r package version 2.1-0.

Li, B. and Wang, S. (2007), "Estimation of subspace arrangements with applications in modeling and segmenting mixed data," *Journal of the American Statistical Association*, 102, 997–1008.

Li, K. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–327.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Nyamundanda, G., Brenna, L., and Gormley, I. C. (2010), "Probabilistic Principal Component Analysis," *BMC Bioinformatics*, 11, 571.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Reich, B. J., Bondall, H. D., and Li, L. (2011), "Sufficient Dimension Reduction via Bayesian Mixture Modeling," *Biometrics*, 67, 886–895.

Schwartz, L. (1965), "On Bayes procedures," *Z. Wahrsch. Verw. Gebiete*, 4, 10–26.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Tipping, M. E. and Bishop, C. M. (1999), "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society Series B*, 61, 611–622.

Titsias, M. K. and Lawrence, N. D. (2010), "Bayesian Gaussian Process Latent Variable Model," in *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics*, eds. Teh, Y. W. and Titterington, D. M., Sardinia Italy, vol. 9, pp. 25–32.

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010), "Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection," *Bayesian Analysis*, 5, 319–344.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer, 4th ed., iSBN 0-387-95457-0. ISBN 0-387-95457-0.

Walker, S. G. (2007), "Sampling the Dirichlet Mixture Model with Slices," *Communications in Statistics; Simulation and Computation*, 36, 45–54.

Wang, H. and Xia, Y. (2008), "Sliced regression for dimension reduction," *Journal of the American Statistical Association*, 103, 811–821.

Wu, Y. and Ghosal, S. (2010), "The L_1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation," *Journal of Multivariate Analysis*, 101, 2411–2419.

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011), "Bayesian Nonparametric Hidden Markov Models with applications in genomics," *Journal of the Royal Statistical Society Series B*, 73, 37–57.

Zhu, Y. and Zeng, P. (2006), "Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression," *Journal of the American Statistical Association*, 86, 1638–1651.