Supervised Dimension Reduction Using Bayesian Mixture Modeling

Kai Mao
Department of Statistical Science
Duke University, NC 27708

Feng Liang
Department of Statistics
University of Illinois at
Urbana-Champaign, IL 61820

Sayan Mukherjee
Departments of Staistical Science
Computer Science and Mathematics
Duke University, NC 27708

Abstract

We develop a Bayesian framework for supervised dimension reduction using a flexible nonparametric Bayesian mixture modeling approach. Our method retrieves the dimension reduction or d.r. subspace by utilizing a dependent Dirichlet process that allows for natural clustering for the data in terms of both the response and predictor variables. Formal probabilistic models with likelihoods and priors are given and efficient posterior sampling of the d.r. subspace can be obtained by a Gibbs sampler. As the posterior draws are linear subspaces which are points on a Grassmann manifold, we output the posterior mean d.r. subspace with respect to geodesics on the Grassmannian. The utility of our approach is illustrated on a set of simulated and real examples. *Some Key Words: supervised dimension reduction, inverse regression, Dirichlet process, factor models, Grassman manifold.*

1 Introduction

Supervised dimension reduction (SDR) or simultaneous dimension reduction and regression can be formulated as finding a low-dimensional subspace or manifold that contains all the predictive information of the response variable. This low-dimensional subspace is often called the dimension reduction (d.r.) space. Projections onto the d.r. space can be used to replace the original predictors, without affecting the prediction. This is a counterpart of unsupervised dimension reduction such as principal components analysis which does not take into account the response variable.

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

$$Y = g(b_1'X, \cdots, b_d'X, \varepsilon) \tag{1}$$

where the column vectors of $B = (b_1, ..., b_d)$ are named the d.r. directions and ε is noise independent of X. In this framework all the predictive information is contained in the d.r. space $\mathcal{B}$ which is the span of the columns of B, since $Y \perp\!\!\!\perp X \mid P_{\mathcal{B}}X$, where $P_{\mathcal{B}}$ denotes the orthogonal projection operator onto the subspace $\mathcal{B}$.

A variety of methods for SDR have been proposed. They can be subdivided into three categories: methods based on gradients of the regression function (Xia et al., 2002; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Mukherjee et al., 2010), methods based on forward regression that investigates the conditional distribution $Y \mid X$ (Friedman and Stuetzle, 1981; Tokdar et al., 2008), and methods based on inverse regression that focuses on $X \mid Y$ (Li, 1991; Cook, 2007; Hastie and Tibshirani, 1996b; Sugiyama, 2007).

In this paper we develop a Bayesian methodology we call Bayesian mixtures of inverse regression (BMI) that extends the model-based approach of Cook (2007). A semiparametric model will be stated. A salient point is that it applies to data generated from distributions where the support of the predictive subspace is not a linear subspace of the predictors but is instead a nonlinear manifold. The projection is still linear but it will contain the nonlinear manifold that is relevant to prediction. A further important point of great interest is that the d.r. subspace is on a so-called Grassmann manifold denoted as $\mathcal{G}_{(d,p)}$ which is defined as the set of all the d dimensional linear subspaces of $\mathbb{R}^p$, and our model allows for rich inference such as uncertainty evaluation by drawing posterior samples (subspaces) from this manifold rather than merely obtaining an optimal point from this manifold as by other SDR methods.

2 Bayesian mixtures of inverse regression

The idea that the conditional distribution of the predictors given the response can provide useful information in the reduction of the dimensions was introduced in sliced inverse

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

regression (SIR) (Li, 1991) for the regression setting and reduced rank linear discriminant analysis for the classification setting. SIR proposes the semiparametric model in (1) and claims that the conditional expectation $E(X \mid Y = y)$, called the inverse regression curve, is contained in the (transformed) d.r. space spanned by the columns of B. SIR is not a model based approach in the sense that a sampling or distributional model is not specified for $X \mid Y$. The idea of specifying a model for $X \mid Y$ is developed in principal fitted component (PFC) models (Cook, 2007). Specifically, the PFC model assumes the following multivariate form for the inverse regression

$$X_y = \mu + A\nu_y + \varepsilon, \quad X_y \equiv X \mid Y = y,$$

where $\mu \in \mathbb{R}^p$ is an intercept; $\varepsilon \sim N(0, \Delta)$ with $\Delta \in \mathbb{R}^{p\times p}$ a random error term; $A \in \mathbb{R}^{p\times d}$ and $\nu_y \in \mathbb{R}^d$ imply that the mean of the (centered) X_y lie in a subspace spanned by the columns of A with ν_y the coordinate (similar to a factor model setting with A the factor loading matrix and ν_y the factor score). Under this model formulation it is important that ν_y is modeled otherwise the above model is an adaptation of principal components regression, see sections 2.1.1 and 2.1.2 for the models used in this paper. In this framework it can be shown $B = \Delta^{-1}A$ (Cook, 2007), so that the columns of $\Delta^{-1}A$ spans the d.r. space.

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

A direct approach to address the first problem is to develop a mixture model, that is, to assume a normal mixture model rather than a simple normal model for X_y. This is the approach taken in mixture discriminant analysis (MDA) (Hastie and Tibshirani, 1994) which utilizes in the classification setting a finite Gaussian mixture model for each class. However MDA can only be applied when the response is discrete rather than continuous, and the prespecification of the (generally unknown) number of mixture components is an issue.

2.1 Model specification

We propose a semiparametric mixture model that generalizes the PFC model

$$X \mid (Y = y, \mu_{yx}, \Delta) \sim N(\mu_{yx}, \Delta) \tag{2}$$

$$\mu_{yx} = \mu + A\nu_{yx} \tag{3}$$

$$\nu_{yx} \sim G_y \tag{4}$$

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

However by considering G_y as a random process hence specifying flexible nonparametric models for $X \mid Y$ we can greatly generalize the PFC model. For example a Dirichlet process prior (DP) (Ferguson, 1973, 1974; Sethuraman, 1994) on G_y leads to a mixture model for $X \mid Y$ due to its discrete property and alleviates the need to prespecify the number of mixture components for $X \mid Y$. For continuous responses the dependent Dirichlet process (DDP) (MacEachern, 1999) or kernel stick-breaking process (Dunson and Park, 2008) is used to model G_y.

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

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

Proof. Assume in the following A and Δ are given. Assume in (3) $\mu = 0$ w.o.l.g. so that $\mu_{yx} = A\nu_{yx}$. Let $p(y|x)$ be the distribution of Y given X. Then

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

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

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

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

$$\begin{aligned}
p(y \mid x) &\propto p(y) \cdot \\
&\int \exp\big(-\frac{1}{2}(P_Ax-\mu_{yx})'\Delta^{-1}(P_Ax-\mu_{yx})\big) d\pi(\mu_{yx})
\end{aligned}$$

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

Given data $\{(x_i, y_i)\}_{i=1}^n$ the following sampling distribution is specified from (2) - (4)

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

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

$$\text{Lik}(\text{data} \mid A, \Delta, \nu, \mu) \propto \det(\Delta^{-1})^{\frac{n}{2}} \cdot \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu - A\nu_i)'\Delta^{-1}(x_i - \mu - A\nu_i)\right] \quad (5)$$

where $\nu = (\nu_1, \cdots, \nu_n)$. To fully specify the model we need to specify the distributions G_{y_i}. The categorical response case is specified in subsection 2.1.1 and the continuous response case is specified in subsection 2.1.2.

2.1.1 Categorical response

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

$$\nu_i \mid (y_i = c) \sim G_c \quad \text{for } c = 1, ..., C, \quad (6)$$

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

$$G_c \sim \text{DP}(\alpha_0, G_0) \quad (7)$$

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

2.1.2 Continuous response

In the case of a continuous response variable it is natural to expect G_{y_1} and G_{y_2} to be dependent if y_1 is close to y_2, that is, we would like to borrow information across the response variables. A natural way of doing this is to use a dependent Dirichlet Process (DDP) prior. The DDP was first introduced in MacEachern (1999) to generate DP to settings where covariates need to be incorporated when modeling a unknown distribution G. Consider the stick breaking construction (Sethuraman, 1994) for $G \sim \text{DP}(\alpha_0, G_0)$

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\nu_h^*}, \quad \nu_h^* \sim G_0$$

where π_h's are weights constructed in a "stick breaking" manner and ν_h^*'s are called "atoms" drawn *a prior* from G_0. Now if G depends on some covariates y and one wants to induce dependence among different such G's through the dependence among different y's one could effectively achieve this by allowing the π_h's, or the ν_h^*'s, or both, to depend on y. For example, one could have $G_y = \sum_{h=1}^{\infty} \pi_{hy}\delta_{\nu_h^*}$ in which now the subscript y is added to G and π_h to show their explicit dependence on y. There are multiple ways to construct such dependent G_y's leading to different DDP (Dunson and Park, 2008; Gelfand et al., 2005; Griffin and Steel, 2006; Iorio et al., 2004; Dunson et al., 2008). In this paper we utilize the kernel stick breaking process (Dunson and Park, 2008) due to its nice properties and computational efficiency. The kernel stick breaking process constructs G_y in such as way that

$$G_y = \sum_{h=1}^{\infty} U(y; V_h, L_h) \prod_{\ell<h}(1 - U(y; V_\ell, L_\ell))\delta_{\nu_h^*} \quad (8)$$
$$U(y; V_h, L_h) = V_h K(y, L_h) \quad (9)$$

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

$$K(y, L_h) = 1_{|y-L_h|<\phi} \quad \text{or} \quad K(y, L_h) = \exp(-\phi|y-L_h|^2). \quad (10)$$

Dependence on the weights $U(y; V_h, L_h)$ in (8) will result in dependence between G_{y_1} and G_{y_2} when y_1 and y_2 are close.

2.2 Inference on the Model Parameters

Given data $\{(x_i, y_i)\}_{i=1}^n$ we would like to infer the model paramaters $A, \Delta, \nu \equiv (\nu_1, \cdots, \nu_n)$. From A and Δ we can compute the d.r. which is the span of $B = \Delta^{-1}A$. The inference will be based on Markov chain Monte Carlo (MCMC) samples from the posterior distribution given the likelihood function in (5) and suitable prior specifications. The inference procedure is a Gibbs sampling scheme which can be broken into four sampling steps: sampling μ, sampling A, sampling Δ^{-1}, and sampling ν. The fourth step will differ based on whether the response variable is continuous or categorical.

Sampling μ and Δ^{-1}

The likelihood function (5) implies a normal distribution in μ and Wishart in Δ^{-1}, so that a noninformative prior on the intercept parameter μ, i.e., $\mu \propto 1$ leads to a normal conditional posterior distribution for μ and a Wishart prior for Δ^{-1} results in a Wishart conditional posterior distribution.

Sampling A

The matrix $A \in \mathbb{R}^{p\times d}$ represents the transformed e.d.r. space and the likelihood (5) implies a normal form in A. We will use the Bayesian factor modeling framework developed in Lopes and West (2004) in which A is viewed as

a factor loading matrix. The key idea is to impose special structure on A to ensure identifiability

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

We specify normal and independent priors for the elements of A

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

the hyper-parameter ϕ_a is specified to take a small value to reflect the vagueness of the prior information. Then conjugacy of the likelihood and the prior leads to a normal conditional posterior for each row of A which we will sequentially update.

Sampling ν for categorical responses

Inference for DP mixture models has been extensively developed in the literature (Escobar and West, 1995; MacEachern and Müller, 1998). We utilize the sampling scheme in Escobar and West (1995) which adopts a marginal approach in sampling from the DP priors. Marginalizing in (6) the unknown distribution G_c leads to the poly-urn representation of the prior for ν_i

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

where G_0 is the base distribution and α_0 is the base concentration parameter. The fact that ν_i should be constrained to have unit variance to ensure identifiability implies that a natural choice of G_0 is $N(0, \mathbf{I}_d)$. Since the likelihood (5) implies a normal form in ν_i, the conditional posterior for ν_i is easy to compute again due to conjugacy (Escobar and West, 1995).

Sampling ν for continuous responses

We follow the sampling scheme for the kernel stick-breaking process developed in Dunson and Park (2008) where sampling details can be referred to. Inference for the DDP is based on a truncation of (8)

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

where H some pre-specified value large integer and $U(y; V_h, L_h) = V_h K(y, L_h) = V_h \exp(-\phi |y - L_h|^2)$ for $h = 1, ..., H-1$ and $U(y; V_H, L_H) = 1$ to ensure that $\sum_{h=1}^{H} U(y; V_h, L_h) \prod_{l<h} (1 - U(y; V_l, L_l)) = 1$. We denote by K_i the cluster label for sample i, that is, $K_i = h$ means that sample i is assigned to cluster h, i.e. $\nu_i = \nu_h^*$. To facilitate sampling V_h we introduce latent variables $Q_{ih} \sim \text{Ber}(V_h)$ and $R_{ih} \sim \text{Ber}(K(y_i, L_h))$ for $i = 1, .., n$ and $h = 1, .., K_i$. Then the iterative sampling procedure among $K_i, \nu_h^*, V_h, Q_{ih}, R_{ih}, L_h$ provides samples of ν_i. Note that priors were previously implied as $\nu_h^* \sim N(0, \mathbf{I}_d)$, $V_h \sim N(v_a, v_b)$, $Q_{ih} \sim \text{Ber}(V_h)$ and $R_{ih} \sim \text{Ber}(K(y_i, L_h))$ and the conditional posteriors are easy to compute due to conjugacy. For L_h we need a Metropolis-Hastings (M-H) sampling step with non-informative prior $L_h \propto 1$ and independent uniform proposal. The kernel precision parameter ϕ in $K(y, L_h) = \exp(-\phi |y - L_h|^2)$ can be pre-specified or sampled. In case of sampling the scheme can be a M-H step with log-normal prior and random walk proposal. For details see Dunson and Park (2008).

2.3 Posterior Inference on the d.r. subspace

Given posterior samples of the parameters A and Δ^{-1} we obtain posterior samples of the d.r. subspace, denoted as $\{\mathcal{B}_1, \cdots, \mathcal{B}_m\}$, where m is the number of the posterior samples. If we fix the dimension d then each subspace is a point on the Grassman manifold denoted as $\mathcal{G}_{(d,p)}$, which is the set of all the d dimensional linear subspaces of $\mathbb{R}^p$. This manifold has a natural Riemannian metric and families of probability distributions can be defined on the Grassmann manifold.

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

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

where the first equation is a singular value decomposition (SVD), $\text{Tr}(\cdot)$ is the matrix trace and $\text{atan}(\cdot)$ is the matrix arctangent. Given the above geodesic distance the mean of the subspaces $\{\mathcal{B}_1, \cdots, \mathcal{B}_m\}$ is the unique subspace with the smallest geodesic distance to the posterior samples

$$\mathcal{B}_{\text{Bayes}} = \arg \min_{\mathcal{B} \in \mathcal{G}_{(d,p)}} \sum_{i=1}^{m} \text{dist}^2(\mathcal{B}_i, \mathcal{B}) \tag{12}$$

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

posterior samples. We obtain a standard deviation estimate of the posterior subspace as

$$\text{std}(\{\mathcal{B}_1, \cdots, \mathcal{B}_m\}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \text{dist}^2(\mathcal{B}_i, \mathcal{B}_{\text{Bayes}})} \quad (13)$$

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

2.4 The $p \gg n$ setting

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

Note in our mixture inverse regression model (2) and (3), μ_{yx} is a mean parameter for $X \mid (Y = y)$, and if $p \gg n$ then it is reasonable to assume that μ_{yx} lies in the subspace spanned by the sample vectors $x_1, \ldots, x_n$ – given the limited sample size constraining the e.d.r. subspace to this subspace is reasonable. By this assumption, $\mu_{yx} - \mu$ and $A\nu_{yx}$, due to equation (3), will also be contained in the subspace spanned by the centered sample vectors. Denote $\tilde{X}$ as the $n \times p$ centered predictor matrix, then a singular value decomposition on $\tilde{X}$ yields $\tilde{X} = U_X D_X V_X'$ with the left eigenvectors $U_X \in \mathbb{R}^{n \times p^*}$ and right eigenvectors $V_X \in \mathbb{R}^{p \times p^*}$ where $p^* \leq n \ll p$. In practice one can select p^* by the decay of the singular values. By the above argument for constraints, we can assume $A = V_X \tilde{A}$ with $\tilde{A} \in \mathbb{R}^{p^* \times d}$. We can also assume that $\Delta = V_X \tilde{\Delta} V_X'$ with $\tilde{\Delta} \in \mathbb{R}^{p^* \times p^*}$. The effective number of parameters is thus hugely reduced.

2.5 Selecting d

In our analysis the dimension of the d.r. subspace d needs to be determined. In a Bayesian paradigm this is formally a model comparison problem involving calculating the Bayes factor which is the ratio of the marginal likelihoods under competing models. The marginal likelihood for a candidate value d is $p(\text{data} \mid d) = \int_\theta p(\text{data} \mid d, \theta) p_{\text{prior}}(\theta) \text{d}\theta$ where θ denotes all the relevant model parameters.

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

3 Application to simulated and real data

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

3.1 Regression on a nonlinear manifold

A popular data set used in the manifold learning literature is the Swiss roll data. We used the following generative model

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

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

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

X_1 and X_3 form an interesting "Swiss roll" shape as illustrated in Figure 1. In this case an efficient dimension reduction method should be able to find the first 3 dimensions.

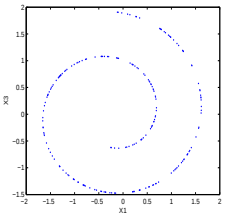

Figure 1: Swiss Roll data. The scatter plot for X_3 v.s. X_1.

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

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

where P_B denotes the orthogonal projection onto the column space of B. For BMI $\hat{B}$ is the posterior Karcher mean as proposed in section 2.3

We did five experiments corresponding to sample size $n = 100, 200, 300, 400, 500$ from the generative model. In each experiment we applied BMI on 10 randomly drawn datasets with sample size n and averaged the accuracies measured as stated above. For BMI we ran 10000 MCMC iterations and used a burn-in of 5000 and set $d = 3$ and used the Gaussian kernel in (10). Figure 2 shows the performance of BMI as well as that by a variety of SDR methods: SIR (Li, 1991), local sliced inverse regression LSIR (Wu et al., 2008), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) and principal Hessian directions (pHd) (Li, 1992). The accuracies for SIR, LSIR, SAVE and PHd are copied from Wu et al. (2008) except for the scenario of $n = 100$. It is clear that BMI consistently has the best accuracy. LSIR is the most competitive of the other methods as one would expect since it shares with BMI the idea of localizing the inverse regression around a mixture or partition.

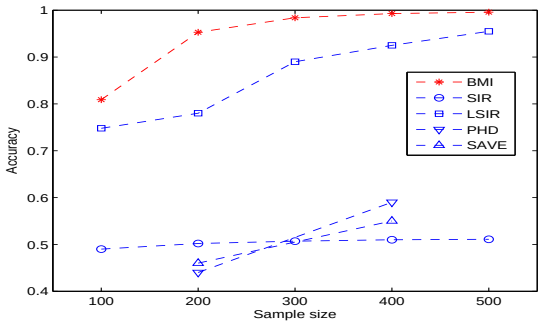

Figure 2: Accuracy for different methods.

Of particular interest is the estimate uncertainty. As stated in section 2.3 the Karcher mean (12) of the posterior samples is taken as a point estimate, and a natural uncertainty measure is simply the standard deviation as defined in (13). For illustration we applied our method on a data set with sample size 400. Figure 3 shows a boxplot for the distances between the posterior sampled subspaces and the posterior Karcher mean subspace and the standard deviation is calculated to be 0.2162. It is also calculated that the distance between the Karcher mean and the true d.r. subspace is 0.2799. It is interesting the see that the true d.r. subspace lies "not far" (compared with the standard deviation) from our point estimate.

We utilized cross-validation to select the number of d.r. directions d in a case of sample size 200. For each value of $d \in \{1, \cdots, 10\}$, we project out-of-sample data onto the d-dimensional space and a nonparametric kernel regression model to predict the response. The error reported is the mean square prediction error. The error v.s. different candidate values of d is depicted in Figure 4. The smallest error corresponds to $d = 3$, the true number of d.r. directions.

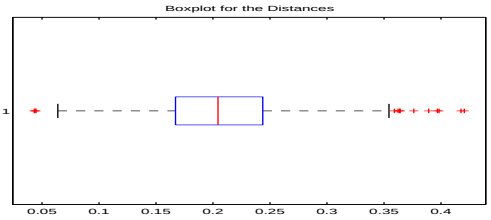

Figure 3: The boxplot for the distances between the posterior samples and their Karcher mean.

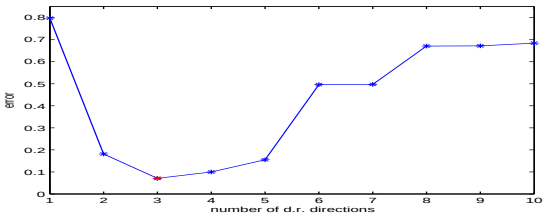

Figure 4: Swiss Roll data: Error v.s. number of d.r directions kept. The minimum one corresponds to $d = 3$, the true value.

3.2 Classification

In Sugiyama (2007) a variety of SDR methods were compared on the Iris data set available from the UCI machine learning repository [1], originally from Fisher (1936). The data consists of 3 classes with 50 instances of each class. Each class refers to a type of Iris plant ("Setosa", "Virginica" and "Versicolour"), and has 4 predictors describing the length and width of the sepal and petal. The methods compared in Sugiyama (2007) were Fisher's linear discriminant analysis (FDA), local Fisher discriminant analysis (LFDA) (Sugiyama, 2007), locality preserving projections (LPP) (He and Niyogi, 2004), LDI (Hastie and Tibshirani, 1996a), neighbourhood component analysis (NCA) (Goldberger et al., 2005), and metric learning by collapsing classes (MCML) (Globerson and Roweis, 2006).

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

In Figures 5 we plot the projection of the data onto a 2 dimensional d.r. subspace. We set $\alpha_0 = 1$ in (7). The classes are separated as are the two clusters in the merged "Setosa", "Virginica" class. Our method is able to further embed the data into a 1 dimensional d.r. subspace while still preserving the separation structure (Figure 6).

Figure 7 is a copy of the Figure 6 in Sugiyama (2007) and provides a comparison of FDA, LFDA, LPP, LDI, NCA, and MCML. Comparing Figure 5 and 6 with Figure 7 we see that BMI and NCA are similar with respect to perfor-

[1] http://archive.ics.uci.edu/ml/datasets/Iris

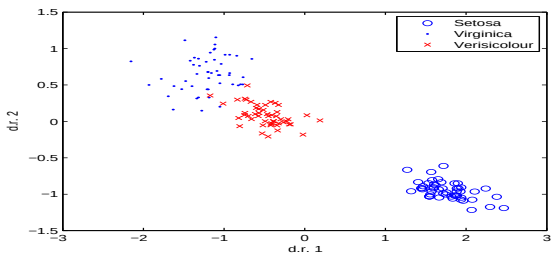

Figure 5: Visualization of the embedded *Iris* data onto a 2 dimension subspace.

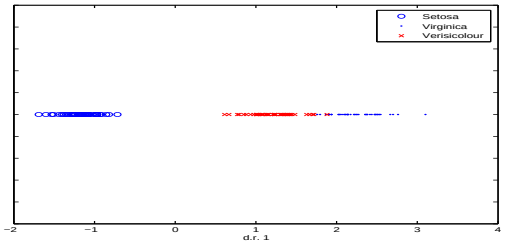

Figure 6: Visualization of the embedded *Iris* data onto a 1 dimensional subspace.

mance, and they both have the advantage of being able to embed this particular data into a 1 dimensional d.r. subspace while the others cannot.

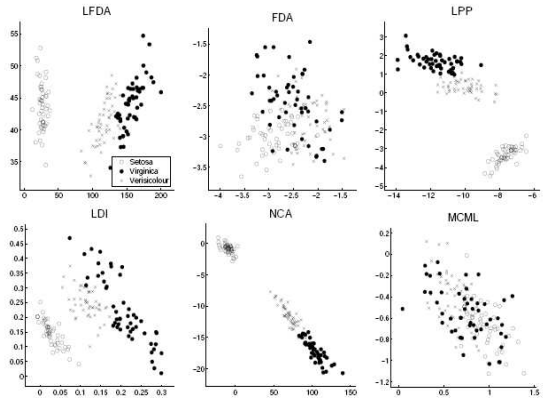

Figure 7: Visualization of the *Iris* data for different methods. (see also Sugiyama (2007) Figure 6.)

3.3 High-dimensional data: digits

The MNIST digits data[2] is commonly used in the machine learning literature to compare algorithms for classification and dimension reduction. The data set consists of $60,000$ images of handwritten digits, $\{0, 1, \ldots, 9\}$ where each image is considered as a vector of $28 \times 28 = 784$ gray-scale pixel intensities. The utility of the digits data is that the d.r. directions have a visually intuitive interpretation.

We apply BMI to two binary classification tasks: digits 3 v.s. 8, and digits 5 v.s. 8. In each task we randomly select 200 images, 100 for each digit. Since the number of predictors is far greater than the sample size ($p \gg n$), we used the modification of BMI described in Section 2.4 and $p^* = 30$ eigenvectors are selected. We run BMI for 10000 iterations with the first as 5000 burn-in and choose $d = 1$. The posterior means of the top d.r. direction, depicted in a 28×28 pixel format, are displayed in Figures 8 and 9. We see that the top d.r. directions precisely capture the difference between digits 3 and 8, an upper left and lower left region, and the different between digits 5 and 8, an upper right and lower left region.

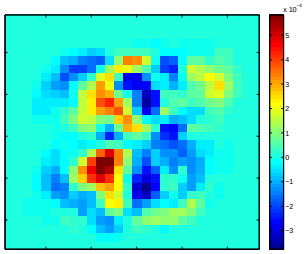

Figure 8: The posterior mean of the top d.r. direction for 3 versus 8, shown in a 28×28 pixel format. Difference between digits is reflected by the red color.

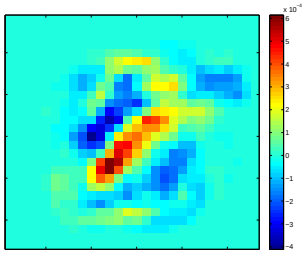

Figure 9: The posterior mean of the top d.r. direction for 5 versus 8, shown in a 28×28 pixel format. Difference between digits is reflected by the red color.

4 Discussion

We have proposed a Bayesian framework for supervised dimension reduction using a highly flexible nonparametric Bayesian mixture modeling approach that allows for natural clustering for the data in terms of both the response and predictor variables. Our model highlights a flexible generalization of the PFC framework to a nonparametric setting and addresses the issue of multiple clusters for a slice of the response. This idea of multiple clusters suggests that this

[2] http://yann.lecun.com/exdb/mnist/

approach is relevant even when the marginal distribution of the predictors is not concentrated on a linear subspace. The idea of modeling nonlinear subspaces is central in the area of manifold learning (Roweis and Saul, 2000; Tenenbaum et al., 2000; Donoho and Grimes, 2003; Belkin and Niyogi, 2004). Our model is one probabilistic formulation of a supervised manifold learning algorithm.

A fundamental issue raised by this methodology is the development of distribution theory on the Grassmann manifold. There has been work on uniform distributions on the Grassmann manifold and we discuss the case corresponding to subspaces drawn from a fixed number of centered normals. To better characterize the uncertainty of our posterior estimates it would be of great interest to develop richer distributions on the Grassmann manifold.

References

Absil, P.-A., R. Mahony, and R. Sepulchre (2004). Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematicae*, 199–220.

Belkin, M. and P. Niyogi (2004). Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning 56*(1-3), 209–239.

Cook, R. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science 22*(1), 1–26.

Cook, R. and S. Weisberg (1991). Discussion of "sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc. 86*, 328–332.

Donoho, D. and C. Grimes (2003). Hessian eigenmaps: new locally linear embedding techniques for high dimensional data. *Proceedings of the National Academy of Sciences 100*, 5591–5596.

Dunson, D. B. and J. Park (2008). Kernel stick-breaking processes. *Biometrika 89*, 268–277.

Dunson, D. B., X. Ya, and C. Lawrence (2008). The matrix stick-breaking process: Flexible bayes meta-analysis. *J. Amer. Statist. Assoc.*, 317–327.

Escobar, M. and M. West (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 577–588.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, 615–629.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annual Eugenics 7*(II), 179–188.

Friedman, J. H. and W. Stuetzle (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, 817–823.

Gelfand, A., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with dirichlet process mixing. *J. Amer. Statist. Assoc.* (471), 1021–1035.

Globerson, A. and S. Roweis (2006). Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems 18*, pp. 451–458.

Goldberger, J., S. Roweis, G. Hinton, and R. Salakhutdinov (2005). Neighbourhood component analysis. In *Advances in Neural Information Processing Systems 17*, pp. 513–520.

Griffin, J. and M. Steel (2006). Order-based dependent dirichlet processes. *J. Amer. Statist. Assoc.*, 179–194.

Hastie and Tibshirani (1994). Discriminant analysis by Gaussian mixtures. *J. Roy.Statist. Soc. Ser. B 58*, 155–176.

Hastie, T. and R. Tibshirani (1996a). Discriminant adaptive nearest neighbor classification. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 607–615.

Hastie, T. and R. Tibshirani (1996b). Discriminant analysis by Gaussian mixtures. *J. Roy.Statist. Soc. Ser. B 58*(1), 155–176.

He, X. and P. Niyogi (2004). Locality preserving projections. In *Advances in Neural Information Processing Systems 16*.

Iorio, M. D., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An anova model for dependent random measures. *J. Amer. Statist. Assoc.*, 205–215.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* (5).

Kendall, W. S. (1990). Probability, convexity and harmonic maps with small image. i. uniqueness and fine existence. *Proc. London Math. Soc.* (2), 371–406.

Li, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc. 86*, 316–342.

Li, K. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Ann. Statist. 97*, 1025–1039.

Lopes, H. F. and M. West (2004). Bayesian model assessment in factor analysis. *statsinica 14*, 41–67.

MacEachern, S. and P. Müller (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics 7*, 223–238.

MacEachern, S. N. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*.

Mukherjee, S. and Q. Wu (2006). Estimation of gradients and coordinate covariation in classification. *J. Mach. Learn. Res. 7*, 2481–2514.

Mukherjee, S., Q. Wu, and D.-X. Zhou (2010). Learning gradients and feature selection on manifolds. *Bernoulli*. in press.

Mukherjee, S. and D. Zhou (2006). Learning coordinate covariances via gradients. *J. Mach. Learn. Res. 7*, 519–549.

Roweis, S. and L. Saul (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science 290*, 2323–2326.

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

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *J. Mach. Learn. Res. 8*, 1027–1061.

Tenenbaum, J., V. de Silva, and J. Langford (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*, 2319–2323.

Tokdar, S., Y. Zhu, and J. Ghosh (2008). A bayesian implementation of sufficient dimension reduction in regression. Technical report, Purdue Univ.

Wu, Q., F. Liang, and S. Mukherjee (2008). Localized sliced inverse regression. Technical report, ISDS, Duke Univ.

Xia, Y., H. Tong, W. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *J. Roy.Statist. Soc. Ser. B 64*(3), 363–410.