

Two models for Bayesian supervised dimension reduction

BY KAI MAO

Department of Statistical Science

Duke University, Durham NC 27708-0251, U.S.A.

km68@stat.duke.edu

QIANG WU

Department of Mathematics

Michigan State University, East Lansing MI 48824, U.S.A.

wuqiang@math.msu.edu

FENG LIANG

Department of Statistics

University of Illinois at Urbana-Champaign, IL 61820, U.S.A.

feng@stat.uiuc.edu

and

SAYAN MUKHERJEE

Department of Statistical Science, Institute for Genome Sciences & Policy, Department of Computer Science

Duke University, Durham NC 27708-0251, U.S.A.

sayan@stat.duke.edu

Abstract

We study and develop two Bayesian frameworks for supervised dimension reduction that apply to non-linear manifolds: Bayesian mixtures of inverse regressions and gradient based methods. Formal probabilistic models with likelihoods and priors are given for both methods and efficient posterior estimates of the effective dimension reduction space and predictive factors can be obtained by a Gibbs sampling procedure. In the case of the gradient based methods estimates of conditional dependence between covariates predictive of the response can also be inferred. Relations to manifold learning and Bayesian factor models are made explicit. The utility of the approach is illustrated on simulated and real examples.

Some Key Words: Supervised dimension reduction, graphical models, manifold learning, inverse regression, factor models

1 Introduction

Simultaneous dimension reduction and regression or supervised dimension reduction (SDR) formulates the problem of high-dimensional regression as finding a low-dimensional subspace or manifold that contains predictive information of the response variable (Li, 1991; Vlassis et al., 2001; Xia et al., 2002; Fukumizu et al., 2003; Li et al., 2004; Goldberger et al., 2005; Fukumizu et al., 2005; Globerson and Roweis, 2006; Martin-Mérino and Róman, 2006; Nilsson et al., 2007; Cook, 2007; Wu et al., 2007; Tokdar et al., 2008; Mukherjee et al., 2009). SDR methods seek to replace the original predictors with projections onto this low dimensional subspace. This low dimensional subspace is often called the effective dimension reduction (e.d.r.) space. 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., 2009), methods based on inverse regression (Li, 1991; Cook and Weisberg, 1991; Hastie and Tibshirani, 1996b; Sugiyama, 2007; Cook, 2007), and methods based on forward regression (Friedman and Stuetzle, 1981; Tokdar et al., 2008).

In this paper we develop Bayesian methodology for two of the approaches to SDR: the first method called Bayesian mixtures of inverse regression (BMI) extends the model-based approach of Cook (2007), the second method Bayesian gradient learning (BAGL) adapts the gradient estimation methods in Xia et al. (2002); Mukherjee and Zhou (2006); Mukherjee and Wu (2006); Mukherjee et al. (2009) to a Bayesian model based setting. In both settings parametric and non-parametric models will be stated or implied. The salient point of both approaches is that they apply to data generated from distributions where the support of the predictive subspace is not a linear subspace of the covariates but is instead a non-linear manifold. The projection is still linear but it will contain the non-linear manifold that is relevant to prediction.

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

$$Y = g(b'_1 X, \dots, b'_d X, \varepsilon) \quad (1)$$

where the span of the vectors $B = (b_1, \dots, b_d)$ is referred to as the effective dimension reduction (e.d.r.) space and ε is noise. One expects $d \ll p$ and the vectors (b_1, \dots, b_d) to form the basis of what is called the central subspace $\mathcal{S} \equiv \mathcal{S}_{Y|X}$ which is defined as the intersection of all subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ having the property that $Y \perp\!\!\!\perp X | P_{\mathcal{S}}X$ where $P_{\mathcal{S}}X$ is the orthogonal projection of X onto \mathcal{S} . In this framework all the predictive information is contained in the central subspace.

Typically the error is assumed to be additive and given a regression function $f(x) = E(Y | X = x)$ the following model holds

$$Y = f(X) + \varepsilon = g(b'_1 X, \dots, b'_d X) + \varepsilon, \quad (2)$$

where again the span of $B = (b_1, \dots, b_d)$ is the e.d.r. space.

There are three approaches to inference of the e.d.r. space

1. Forward regression: The conditional probability $Y | X$ is directly modeled. A classic example of this approach is Projection Pursuit Regression (PPR) (Friedman and Stuetzle, 1981). A modern Bayesian example is proposed in Tokdar et al. (2008) where a variant of logistic Gaussian processes is utilized to model $Y | P_{\mathcal{S}}X$.
2. Inverse regression: The conditional distribution $X | Y$ is the focus of these approaches. The classic example is sliced inverse regression (SIR) (Li, 1991) which estimates the inverse regression curve $E(X | Y)$ to infer about the e.d.r. space. This approach is extended to more general settings in Principal Fitted Component (PFC) models (Cook, 2007).
3. Learning gradients: The observation that the gradient of the regression function, ∇f , lies in the e.d.r. space motivates this approach. A variety of methods exist for inference of the gradient (Xia et al., 2002; Mukherjee and Zhou, 2006; Mukherjee and Wu, 2006; Mukherjee et al., 2009; Wu et al., 2007).

The Bayesian formulation of the inverse regression framework has a natural model-based underpinning based on distribution theory. In contrast, the Bayesian formulation of the gradient based framework will highlight geometric considerations and provide a rigorous link to manifold learning.

2 Bayesian mixtures of inverse regression (BMI)

The idea that the conditional distribution of the predictor given the response can provide useful information in the reduction of the dimensions was introduced in sliced inverse regression (SIR) (Li, 1991) for the regression setting and reduced rank linear discriminant analysis for the classification setting. SIR proposes the semiparametric model in (1) and claims that the conditional expectation $E(X | Y = y)$, called the inverse regression curve, is contained in the (transformed) e.d.r. space B . SIR is not a model based approach in the sense that a sampling or distributional model is not specified for $X | Y$. The idea of specifying a model for $X | Y$ is developed in principal fitted component (PFC) models (Cook, 2007). Specifically, the PFC model assumes the following multivariate form for the inverse regression

$$X_y = \mu + A\nu_y + \varepsilon \quad (3)$$

where $X_y \equiv X | Y = y$; $\mu \in \mathbb{R}^p$ is an intercept; $\varepsilon \sim N(0, \Delta)$ with $\Delta \in \mathbb{R}^{p \times p}$ is a random error term; $A \in \mathbb{R}^{p \times d}$, $\nu_y \in \mathbb{R}^d$ imply that the mean of the (centered) X_y lie in a subspace spanned by the column 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). In this framework it can be shown $B = \Delta^{-1}A$ (Cook, 2007), so that the columns of $\Delta^{-1}A$ spans the e.d.r. space.

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

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.

2.1 Model specification

We propose a semiparametric mixture model that generalizes the PFC model (3):

$$X | (Y = y, \mu_{yx}, \Delta) \sim N(\mu_{yx}, \Delta) \quad (4)$$

$$\mu_{yx} = \mu + A\nu_{yx} \quad (5)$$

$$\nu_{yx} \sim G_y \quad (6)$$

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 (3); $\nu_{yx} \in \mathbb{R}^d$ is analogous to ν_y in (3) except it depends on y and the marginal distribution of X , and it follows a distribution G_y that depends on y . Note that (3) can be recovered by assuming $G_y = \delta_{\nu_y}$ is a point mass at ν_y , in this case $\nu_{yx} \equiv \nu_y$.

However by considering G_y as a random process hence specifying flexible non-parametric models for $X | Y$ we can greatly generalize (3). For example a Dirichlet process prior (DP) (Ferguson, 1973, 1974; Sethuraman, 1994) on G_y leads to a mixture model for $X | Y$ due to its discrete property and alleviates the need to prespecify the number

of mixtures for $X | Y$. In the setting of a continuous response the dependent Dirichlet process (MacEachern, 1999; Dunson and Park, 2008) can be used to allow dependence between G_y 's.

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

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

Proof. Assume in the following distributions A and Δ are implicitly given. Assume in (5) $\mu = 0$ w.o.l.g. Let $p(y|x)$ be the distribution of Y given X . Then

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

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

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

So that for any y_1 and y_2 ,

$$\frac{p(y_1 | x)}{p(y_2 | x)} = \frac{p(y_1) \int \exp\left(-\frac{1}{2}(P_A x - \mu_{y_1 x})' \Delta^{-1}(P_A x - \mu_{y_1 x})\right) d\pi(\mu_{y_1 x})}{p(y_2) \int \exp\left(-\frac{1}{2}(P_A x - \mu_{y_2 x})' \Delta^{-1}(P_A x - \mu_{y_2 x})\right) d\pi(\mu_{y_2 x})}$$

only depends on $P_A x$, 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 (4) - (6)

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

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

$$\text{Lik}(\text{data} | A, \Delta, \nu, \mu) \propto \det(\Delta^{-1})^{\frac{n}{2}} \exp\left[-\frac{1}{2} \sum_{i=1}^n (x_i - \mu - A\nu_i)' \Delta^{-1} (x_i - \mu - A\nu_i)\right], \quad (7)$$

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

2.1.1 Categorical response

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

$$\nu_i | (y_i = c) \sim G_c \quad \text{for } c = 1, \dots, C, \quad (8)$$

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

$$G_c \sim \text{DP}(\alpha_0, G_0) \quad (9)$$

with α_0 is a weight 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 $X | Y$. 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 . We would like to borrow information across the response variables. A natural way of doing this is to use a dependent DP prior, specifically the kernel stick-breaking process (Dunson and Park, 2008)

$$G_y = \sum_{h=1}^{\infty} U(y; V_h, L_h) \prod_{\ell < h} (1 - U(y; V_\ell, L_\ell)) \delta_{\nu_h^*} \quad (10)$$

$$U(y; V_h, L_h) = V_h K(y, L_h) \quad (11)$$

where L_h is a random location in the domain of y , $V_h \sim \text{Be}(v_a, v_b)$ 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

$$K(y, L_h) = 1_{|y - L_h| < \phi} \quad \text{or} \quad K(y, L_h) = \exp(-\phi |y - L_h|^2). \quad (12)$$

Dependence on the weights $U(y; V_h, L_h)$ in (10) will result in dependence between G_{y_1} and G_{y_2} when y_1 and y_2 are close.

2.2 Inference

Given data $\{(x_i, y_i)\}_{i=1}^n$ we would like to infer the model parameters $A, \Delta, \nu \equiv (\nu_1, \dots, \nu_n)$. From A and Δ we can compute the e.d.r. which is the span of $B = \Delta^{-1}A$. The inference will be based on Markov chain Monte Carlo samples from the posterior distribution given the likelihood function in (7) and suitable prior specifications. The inference procedure can be broken into four sampling steps: sampling μ , sampling A , sampling Δ^{-1} , and sampling ν . The fourth step will differ based on whether the response variable is continuous or categorical.

Sampling μ

A noninformative prior on the intercept parameter μ , i.e., $\mu \propto 1$, combined with the likelihood function (7), leads to normal full conditional posterior distribution

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

Sampling A

The matrix $A \in \mathbb{R}^{p \times d}$ represents the transformed e.d.r. space and the likelihood (7) implies that A is normally distributed. We will use the Bayesian factor modeling framework developed in Lopes and West (2004) to model A . In this framework A is viewed as a factor loading matrix. The key idea in (Lopes and West, 2004) is to impose special structure on A to ensure identifiability

$$A = \begin{pmatrix} a_{11} & 0 & 0 \\ \vdots & \ddots & 0 \\ a_{d1} & \dots & a_{dd} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pd} \end{pmatrix} \quad (13)$$

where $\{a_{ii}\}_{i=1}^d$ are further constrained to be positive.

We specify normal and independent priors for the elements of A

$$\begin{aligned} a_{\ell j} &\sim N(0, \phi_a^{-1}), \ell > j, \ell = 2, \dots, p \\ a_{\ell \ell} &\sim N(0, \phi_a^{-1})1_{(a_{\ell \ell} > 0)}, \ell = 1, \dots, d, \end{aligned}$$

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

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

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

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

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

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

Sampling Δ

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

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

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 (8) 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 scale parameter. The fact that ν_i should be constrained to have unit variance to ensure identifiability implies that a natural choice of G_0 is $N(0, \mathbf{I}_d)$. Combining with the likelihood (7) the full conditional for ν_i is

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

where

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

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

Sampling ν for continuous responses

We follow the sampling scheme for the kernel stick-breaking process developed in Dunson and Park (2008). Inference for the DDP is based on a truncation of (10)

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

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

The following iterative procedure provides samples of ν_i

1. Sample the cluster membership K_i

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

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

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

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

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

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

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

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

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

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

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

A Metropolis-Hastings step is taken with the proposal distribution $L_h^* \sim \text{Unif}(\min_{i \in \{R_{ih}=1\}}(y_i), \max_{i \in \{R_{ih}=1\}}(y_i))$ and the acceptance ratio is calculated to be

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

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

Let $\tilde{\phi} = \log(\phi)$. We place a normal prior on $\tilde{\phi}$, namely, $\tilde{\phi} \sim N(\mu_{\tilde{\phi}}, \sigma_{\tilde{\phi}}^2)$ which implies a log-normal prior on ϕ , and a proposal distribution a random walk $\tilde{\phi}^* \sim N(\tilde{\phi}, \sigma_{\text{prop}}^2)$, with $(\mu_{\tilde{\phi}}, \sigma_{\tilde{\phi}}^2, \sigma_{\text{prop}}^2)$ pre-specified hyper-parameters. The acceptance ratio is thus

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

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

2.3 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 (4) and (5), μ_{yx} is a mean parameter for $X | (Y = y)$, and if $p \gg n$ then it is reasonable to assume that μ_{yx} lies in the subspace spanned by the sample vectors x_1, \dots, x_n – given the limited sample size constraining the e.d.r. subspace to this subspace is reasonable. By this assumption, $\mu_{yx} - \mu$ and $A\nu_{yx}$, due to equation (5), 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.

2.4 Selecting d

In our analysis the dimension of the e.d.r. subspace d needs to be determined. In a Bayesian paradigm this is formally a model comparison problem and for two candidate values d_1 and d_2 the Bayes factor can be used for model selection

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

with the marginal likelihood

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

where θ denotes all the relevant model parameters.

The marginal likelihood in our case is obviously not analytically available. Various approximation methods are listed in Lopes and West (2004) yet none of them prove to be computationally efficient in our case. 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.

2.5 Model comments

The Bayesian mixtures of inverse regression (BMI) model highlights a simple flexible generalization of the PFC framework to a non-parametric setting and addresses the issue of multiple clusters for a slice of the response. This idea of multiple clusters suggests that this approach is relevant even when the marginal distribution of the covariates is not concentrated on a linear subspace. The idea of modeling non-linear 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). BMI is one probabilistic formulation of a supervised manifold learning algorithm. The connection between SDR and manifold learning will be made explicit in the next section.

3 Bayesian gradient learning (BAGL)

The gradient framework for SDR highlights the geometry of the marginal distribution of the covariates and the geometric properties of the regression function. Recall that under the semi-parametric model in (2)

$$Y = f(X) + \varepsilon = g(b'_1 X, \dots, b'_d X) + \varepsilon.$$

If the error term is i.i.d. then data are drawn i.i.d. from a joint distribution on $\rho(X, Y)$. We assume that the marginal distribution ρ_X is concentrated on a manifold $\mathcal{M} \subset \mathbb{R}^p$ of dimension $d_{\mathcal{M}} \ll p$.

We first state the relevant mathematical concepts and quantities that we use in our formulation. To fix mathematical ideas we assume the existence of an isometric embedding from the manifold to the ambient space, $\varphi : \mathcal{M} \rightarrow \mathbb{R}^p$. The observed explanatory variables are the image of points drawn from a distribution concentrated on the manifold, $x_i = \varphi(q_i)$ where $(q_i)_{i=1}^n$ are points concentrated on the manifold. The gradient on the manifold $\nabla_{\mathcal{M}} f$ is a well defined mathematical quantity and is a $d_{\mathcal{M}}$ -dimensional vector. However, since we only obtain observations in the p -dimensional ambient space we cannot compute the gradient on the manifold. Given data one can estimate the gradient in the ambient space, \vec{f} , which is a p -dimensional vector. It was shown in Mukherjee et al. (2009) that under weak conditions on the manifold and regression an estimate of the gradient in the ambient space, \vec{f} , is consistent in the following sense

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

where $(d\varphi)^*$ is the dual of $d\varphi$. This suggests that the gradient estimate in the ambient space provides information on the gradient on the manifold, even if the gradient on the ambient space $\nabla f = \left(\frac{\partial f}{\partial x^1}, \dots, \frac{\partial f}{\partial x^p} \right)'$ is not well defined.

Assuming that the gradient in the ambient space is well defined the gradient outer product (GOP) matrix is

$$\Gamma = E_X [(\nabla f) (\nabla f)']. \quad (14)$$

A key observation in Wu et al. (2007) is that under the model specified by equation (2) the eigenvectors $\{v_1, \dots, v_d\}$ corresponding to the largest d eigenvalues of Γ span the e.d.r. space

$$\text{span}(b_1, \dots, b_d) = \text{span}(v_1, \dots, v_d)$$

This observation formalized in (Wu et al., 2007) is the key link between SDR and learning gradients on manifolds. If the gradient in the ambient space is not well defined then the result holds asymptotically for an empirical estimate of the GOP

$$\hat{\Gamma} = (\vec{f}) (\vec{f})'.$$

The GOP matrix can also be defined in terms of statistical quantities of the generative model underlying the data (Wu et al., 2007). In the setting of linear regression

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

where $\Omega_{X|Y} = \text{cov}[E(X|Y)]$ is the covariance of the inverse regression, σ_Y^2 is the variance of the response variable, σ_ε^2 is the variance of the error, and $\Sigma_X = \text{cov}(X)$ is the covariance of the explanatory variables. A similar result holds for non-linear functions that are smooth in this case assuming there exists \mathcal{I} partitions R_i of the explanatory variables such that

$$f(x) = \beta'_i x + \varepsilon_i, \quad E\varepsilon_i = 0 \quad \text{for } x \in R_i, \quad (15)$$

then

$$\Gamma = \sum_{i=1}^{\mathcal{I}} \rho(R_i) \sigma_i^2 \left(1 - \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2}\right)^2 \Sigma_i^{-1} \Omega_i \Sigma_i^{-1}, \quad (16)$$

where $\Sigma_i = \text{cov}(X \in R_i)$ is the covariance matrix of the explanatory variables in partition R_i , $\sigma_i^2 = \text{var}(Y|X \in R_i)$ is the variance of the response variables in partition R_i , $\Omega_i = \text{cov}[E(X \in R_i|Y)]$ is the the covariance of the inverse regression in partition R_i , and $\rho(R_i)$ is the measure of partition R_i with respect to the marginal distribution. This averaging over partitions or mixtures is in the same spirit as decomposing the mixtures of normals proposed in the BMI model.

This interpretation of the GOP as a covariance matrix allows for the inference of conditional dependence between covariates given the response. The theory of Gauss-Markov graphs (Speed and Kiiveri, 1986; Lauritzen, 1996) can be applied to the GOP. If we consider the GOP matrix Γ as a covariance matrix as defined above then the precision matrix $J = \Gamma^{-1}$ provides an estimate of the conditional dependence of the explanatory variables with respect to the response variable. The modeling assumption of this construction is that the matrix J is sparse with respect to the factors or directions (b'_1, \dots, b'_d) rather than the p explanatory variables. Given J the partial correlation matrix R is defined as

$$R_{ij} = -\frac{J_{ij}}{\sqrt{J_{ii}J_{jj}}}.$$

3.1 Model specification

Given data $\{(x_i, y_i)\}_{i=1}^n$ we specify the following regression model

$$y_i = \frac{1}{n} \left[\sum_{j=1}^n f(x_j) + \vec{f}(x_i)'(x_i - x_j) + \varepsilon_{ij} \right], \quad \text{for } i = 1, \dots, n, \quad (17)$$

$$\varepsilon_{ij} = y_i - f(x_j) - \vec{f}(x_i)'(x_i - x_j), \quad \text{for } i, j = 1, \dots, n \quad (18)$$

where f models the regression function, \vec{f} models the gradient, and ε_{ij} has both stochastic and deterministic components varying monotonically as a function of the distance between two observations x_i and x_j . This model is obtained by coupling the first order Taylor series expansion of the regression function $f(x)$ around a point u

$$f(x) = f(u) + \nabla f(x)'(x - u) + \varepsilon_d, \quad (19)$$

where the deterministic error term $\varepsilon_d = O(\|x - u\|^2)$ is a function of the distance between x and u with the regression model specified in (2) with additive independent stochastic noise. Specifying a parametric or non-parametric model for f and \vec{f} in addition to a model for ε_{ij} will define a likelihood model.

We model ε_{ij} as a random quantity and use a very simple spatial model to specify the covariance structure. We first define an association matrix with $w_{ij} = \exp(-\|x_i - x_j\|^2/2s^2)$ with fixed bandwidth parameter s . We then define $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, (\phi/w_{ij})^{-1})$ where ϕ will be a random scale parameter. Define the vector $\varepsilon_{i\bullet} = (\varepsilon_{i1}, \dots, \varepsilon_{in})'$, a joint probability density function on this vector can be used to specify a likelihood function for the data. We specify the following model for $\varepsilon_{i\bullet}$

$$p(\varepsilon_{i\bullet}) \propto \phi^{\frac{n}{2}} \exp \left\{ -\frac{\phi}{2} (\varepsilon'_{i\bullet} W_i \varepsilon'_{i\bullet}) \right\}, \quad (20)$$

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

We use a kernel model for the regression function and gradient

$$f(x) = \sum_{i=1}^n \alpha_i K(x, x_i), \quad \vec{f}(x) = \sum_{i=1}^n \mathbf{c}_i K(x, x_i) \quad (21)$$

where $\alpha = (\alpha_1, \dots, \alpha_n)' \in \mathbb{R}^n$, $C = (\mathbf{c}_1, \dots, \mathbf{c}_n) \in \mathbb{R}^{p \times n}$. The use of these kernel estimators for Bayesian regression models was justified in Pillai et al. (2007). Substituting the above representation in equation (18) results in the following parametrized model

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

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

$$y_i \mathbf{1} = K \alpha + D_i C K_i + \varepsilon_{i\bullet} \quad (22)$$

where $\mathbf{1}$ is the $n \times 1$ vector of all 1's, K_i is the i -th column of the gram matrix K where $K_{ij} = k(x_i, x_j)$, E is the $n \times p$ data matrix and $D_i = \mathbf{1}x'_i - E$. The matrix the number of parameters in the matrix C can be greatly reduced due to the fact that the linearization imposed by the first order Taylor series expansion in (17) imposes the constraint that the gradient estimate must be in the span of differences between data points, $M_X = (x_1 - x_n, x_2 - x_n, \dots, x_{n-1} - x_n) \in \mathbb{R}^{p \times n}$. The rank of this matrix is $\tilde{d} \leq \min((n-1), p)$ and the singular value decomposition yields $M_X = V_M \Lambda_M U'_M$ where $V_M = (v_1, \dots, v_{n-1})$ and U_M are the left and right eigenvectors and Λ_M is a matrix of the singular values. For a fixed d^* corresponding to large singular values we select the corresponding left eigenvectors $\tilde{V} = (v_1, \dots, v_{d^*})$ and define a new set of parameters $\tilde{C} = \tilde{V}' C$ and the define $D_i = \tilde{D}_i \tilde{V}'$. A spectral decomposition can also be applied to the gram matrix K resulting in $K = F \Lambda_K F'$. Note that $K \alpha = F \beta$ where $\beta = \Lambda_K F' \alpha$. We can again select columns of F corresponding to the m largest eigenvalues. Given the above re-parametrization we have for the i -th observation

$$y_i \mathbf{1} = F \beta + \tilde{D}_i \tilde{C} K_i + \varepsilon_{i\bullet} \quad (23)$$

Given the probability model for the error vector in (20), the likelihood of our model given observations data = $\{(x_1, y_1), \dots, (x_n, y_n)\}$ is

$$\text{Lik}(\text{data} \mid \phi, \beta, \tilde{C}) \propto \phi^{\frac{n^2}{2}} \exp \left\{ -\frac{\phi}{2} \sum_{i=1}^n \left(y_i \mathbf{1} - F \beta - \tilde{D}_i \tilde{C} K_i \right)' W_i \left(y_i \mathbf{1} - F \beta - \tilde{D}_i \tilde{C} K_i \right) \right\}, \quad (24)$$

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

3.2 Inference

Given data $\{(x_i, y_i)\}_{i=1}^n$ we would like to infer the GOP Γ which we can use to estimate the e.d.r. as well as the conditional independence matrix J . The inference will be based on Markov chain Monte Carlo samples from the posterior distribution given the likelihood function in (24) and suitable prior specifications. We first provide the procedure for a continuous response and then the setting of a binary response.

Continuous response

We adapt the prior specification developed in West (2003) for factor models in high dimensional regression to provide an efficient sampling scheme for posterior inference of the model parameters. Normal priors are placed on the

elements of β as well as the columns of \tilde{C} without losing the covariance structure in the data due to the orthogonality of the columns of F and \tilde{D}_i in (23). The prior specification for these parameters are

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

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

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

where \tilde{C}_j is the j -th column of \tilde{C} and $a_\psi, b_\psi, a_\varphi, b_\varphi, \omega$ are pre-specified hyper-parameters, and an improper prior for ϕ is used. These independent priors induce generalized g-priors for α and C in (22).

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

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

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

where β_i is the i -th element of β ;

2. Update Δ_φ :

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

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

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

3. Update β :

$$\beta \mid \text{data}, \tilde{C}, \Delta_\psi, \phi \sim N(\mu_\beta, \Sigma_\beta)$$

with

$$\Sigma_\beta = \left(F' \left(\sum_{i=1}^n \phi W_i \right) F + \Delta_\psi \right)^{-1},$$

$$\mu_\beta = \phi \Sigma_\beta F' \sum_{i=1}^n W_i (y_i \mathbf{1} - \tilde{D}_i \tilde{C} K_i);$$

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

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

$$\tilde{C}_j \mid \text{data}, \tilde{C}_{\setminus j}, \Delta_\psi, \phi \sim N(\mu_j, \Sigma_j),$$

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

$$b_{ij} = y_i \mathbf{1} - F \beta - \tilde{D}_i \sum_{k \neq j} \tilde{C}_k K_{ik}$$

$$\Sigma_{j0} = \left(\phi \sum_{i=1}^n K_{ij}^2 \tilde{D}_i' W_i \tilde{D}_i \right)^{-1},$$

$$\mu_{j0} = \phi \Sigma_j \sum_{i=1}^n K_{ij} \tilde{D}_i' W_i b_{ij}$$

$$\Sigma_j = (\Sigma_{j0}^{-1} + \Delta_\varphi)^{-1},$$

$$\mu_j = \Sigma_j (\Sigma_{j0}^{-1} \mu_{j0}).$$

5. Update ϕ :

$$\phi \mid \text{data}, \tilde{C}, \beta \sim \text{Gamma}(a, b),$$

where

$$a = \frac{n^2}{2},$$

$$b = \frac{1}{2} \left(\sum_{i=1}^n \left[y_i \mathbf{1} - F\beta - \tilde{D}_i \tilde{C} K_i \right]' W_i \left[y_i \mathbf{1} - F\beta - \tilde{D}_i \tilde{C} K_i \right] \right).$$

Given draws $\{\tilde{C}^{(t)}\}_{t=1}^T$ from the posterior we can compute $\{C^{(t)}\}_{t=1}^T$ from the relation $\tilde{C} = \tilde{V}'C$. which allows use to compute a gradient outer product for each draw

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

Given these instances of the gradient outer product we can compute the posterior mean gradient outer product matrix as well as its variance

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

An eigen-decomposition of $\hat{\mu}_{\Gamma, D}$ provides us with an estimate of the e.d.r. space \hat{B} and $\hat{\sigma}_{\Gamma, D}$ provides us with an estimate of the uncertainty of stability in our estimate of the e.d.r. For inference of conditional independence we first compute the conditional independence and partial correlation matrices

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

using a pseudo-inverse to compute $(\Gamma_D^t)^{-1}$. The mean and variance of the posterior estimates of conditional independence as well as partial correlations can be computed as above using $\{J^{(t)}\}_{t=1}^T$ and $\{R^{(t)}\}_{t=1}^T$

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

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

Binary response

The extension to classification problems where responses are $y_i = 1/0$ using a probit link function is implemented using a set of latent variables $Z = (z_1, \dots, z_n)'$ modeled as a truncated normal distribution with standard variance. In this setting $\phi \equiv 1$ and almost the same Gibbs sampler as above applies with all $\phi \equiv 1$ and y_i replaced by z_i and a step added to sample the latent variable:

Update Z :

For $i = 1, \dots, n$

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

where N^+ (N^-) denotes the positive (negative) truncated normal distributions and $(\eta_1, \dots, \eta_n)' = F\beta$.

3.3 Selecting d

The decision of how many of the e.d.r. dimensions to keep can in theory rely upon the posterior distribution of the eigenvalues of the gradient outer product matrices drawn by simulating from the posterior. In practice we used the cross-validation procedure outlined in the BMI case.

3.4 Modeling comments

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

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

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

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

4 Application to simulated and real data

To illustrate the efficacy of BMI and BAGL we apply them to simulated and real data. The first simulation illustrates how these two methods capture information on non-linear manifolds. The second data set is used to compare these two methods to a variety other supervised dimension reduction methods in the classification setting. The third data set illustrates that the methods can be used in high-dimensional data. The fourth example highlights the graphical modeling or inference of conditional independence using real and simulated data.

4.1 Regression on a non-linear manifold

A popular data set used in the manifold learning literature is the swiss roll data, see Figure 1. We used the following generative model

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

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

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

X_1 and X_3 form an interesting ‘‘Swiss roll’’ shape as illustrated in Figure 1(b) and the nonlinear relationship between Y and X_1, X_2, X_3 is illustrated in Figure 1 (a). In this case an efficient dimension reduction method should be able to find the first 3 dimensions.

We used the following metric proposed in Wu et al. (2008) to measure the accuracy in estimating the e.d.r. space. Let $\hat{B} = (\hat{\beta}_1, \dots, \hat{\beta}_d)$ denote the estimate of B , 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 and BAGL \hat{B} are the orthonormalized posterior mean of B .

We draw five data sets with sample size $n = 200, 300, 400, 500, 600$ from the generative model. We run BAGL and BMI on these data sets and compare their performance to a variety of SDR methods: SIR (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), principal Hessian directions (pHd) (Li, 1992), and local sliced inverse regression LSIR (Wu et al., 2008). For BMI and BAGL we ran 10000 MCMC iterations and used a burn-in of 5000. For BMI we set $d = 3$ and for BAGL we used the Gaussian kernel in (12) with fixed precision parameter $\phi = 4$ and set $m = d^* = p = 10$.

The accuracy for the various methods using the above defined metric is shown in Figure 2. BAGL and BMI have the best accuracy, especially when sample size is small. 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.

To illustrate that BMI is borrowing of information across the response variables we plot in Figure 3 the cluster labels for all the samples as ordered in terms of the magnitude of response for the last MCMC iteration. Samples with similar responses tend to be clustered and have similar underlying clustering distributions which change "gradually" with increasing response instead of "rigidly" as what would be obtained by the slicing procedure in SIR.

Figure 4 shows for the BMI and BAGL the posterior mean and 90% credible intervals for the estimated 3 e.d.r. directions for a sample size $n = 200$. The noise dimensions have posterior mean at nearly 0 with relatively high certainty indicated by the credible intervals. The first 3 true signal dimensions, on the other hand, are clearly verified.

We utilized cross-validation to select the number of e.d.r. directions d for both BAGL and BMI in the case of sample size $n = 200$. We used, for each value of $d \in \{1, \dots, 10\}$, BMI and BAGL to project out-of-sample data onto the d -dimensional space and a non-parametric 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 5. For both BAGL and BMI the smallest error corresponds to $d = 3$, the true number of e.d.r. directions.

4.2 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 and BAGL can find multiple clusters we merge "Setosa", "Virginica" into a single class and examine whether we are able to separate them.

In Figures 6 (a) and (b) we plot the projection of the data onto a 2 dimensional e.d.r. subspace by BMI and BAGL, respectively. For BMI we set $\alpha_0 = 1$ in (9) and for BAGL $m = d^* = p = 4$. For both methods the classes are separated as are the two clusters in the merged "Setosa", "Virginica" class. BMI is able to further embed the data into a 1 dimensional e.d.r. subspace while still preserving the separation structure (Figure 6 (c)). Figure 7 is a copy of Figure 6 in Sugiyama (2007) and provides a comparison of FDA, LFDA, LPP, LDI, NCA, and MCML. Comparing Figure 6 with Figure 7 we see that BMI and NCA are similar with respect to performance and BAGL and LPP are similar. Both BAGL and LPP highlight the manifold perspective for dimension reduction.

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

4.3 High-dimensional data: digits

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

We apply both BMI and BAGL 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 2.3 and $p^* = 30$ eigenvectors are selected. We run both BMI and BMAGL for 10000 iterations with the first as 5000 burn-in. We choose $d = 1$. The posterior means of the top e.d.r. direction, depicted in a 28×28 pixel format, are displayed in Figures 8 and 9. We see that the top e.d.r. directions precisely capture the difference between digits 3 and 8, an upper left and a lower left region, and the difference between digits 5 and 8, an upper right and lower left region.

4.4 Inference of conditional dependence

We illustrate how BAGL can be used to infer conditional dependence of variables relevant in predicting the response.

The following linear regression model is specified

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

where $\theta \sim N(0, 1)$ and

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

where $\varepsilon \sim N(0, 0.25)$. One hundred samples were drawn from this model and we estimated the mean and standard deviation of the gradient outer product matrix, see Figure 10 (a) and (b). The partial correlation matrix and its standard deviation are also displayed in Figure 10(c) and (d). The inference consistent with the estimate of the partial correlation structure is that X_1, X_3, X_5 are negatively correlated with respect to variation in the response and X_2 and X_4 are not correlated with respect to variation in the response. This relation is displayed in the graphical model in Figure 11(b). The graphical model corresponding to the covariance of the explanatory variables alone is displayed in Figure 11(a).

4.4.1 Tumorigenesis example

We further consider a practical problem in cancer genetics, modeling tumorigenesis. Genetic models of cancer progression are of great interest to better understand the initiation of cancer as well as the progression of disease into metastatic states. In Edelman et al. (2008) models of tumor progression in prostate cancer as well as melanoma were developed. Instead of using gene expression measurements as the predictor variables a summary statistic that assayed the differential enrichment of a priori defined sets of genes in individual samples (Edelman et al., 2006, 2008) was used. These a priori defined gene sets correspond to genes known to be in signalling pathways or have functional relations.

This idea was applied to a data set consisting of 22 benign prostate samples and 32 malignant prostate samples (Tomlins et al., 2007). The 20,000 dimensional expression profile for each sample was reduced to a space of differential enrichment of 522 gene sets or pathways (Edelman et al., 2008) for each sample. This is a classification problem to which we applied BAGL to infer a mean posterior conditional independence matrix as well as the uncertainty in dependence inferences. For visualization purposes we focus on the 16 pathways most relevant with respect to predicting

²<http://yann.lecun.com/exdb/mnist/>

progression, the 16 pathways corresponding to predictors for which the partial derivative is large. For these gene sets we plot the conditional independence matrix and the variance of the elements in the matrix in Figure 12. Red edges correspond to positive partial correlations and blue for negative. The width of the edges correspond to the degree of uncertainty, edges we are more sure of are thicker. This graph offers a great deal of interesting biology to explore some of which is known, see Edelman et al. (2008) for details on the biology.

5 Discussion

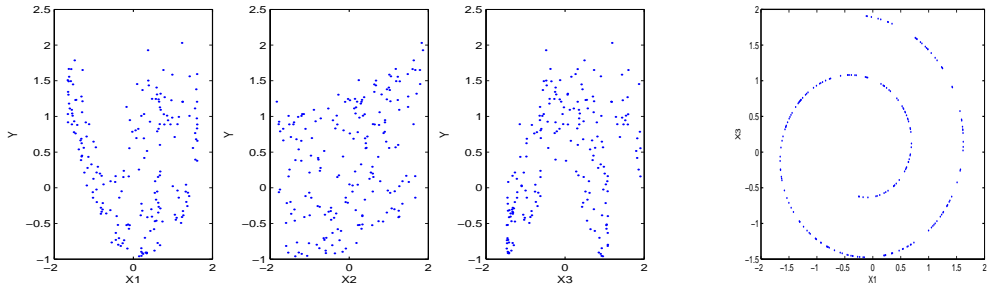
We develop two Bayesian models for SDR. The first model BMI extends model based approaches for inverse regression to the mixture models. The second model BAGL starts from a manifold perspective to develop a probabilistic model. Both approaches apply to the setting where the marginal distribution of the predictors lie on a manifold.

References

- 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.
- Edelman, E., J. Guinney, J.-T. Chi, P. Febbo, and S. Mukherjee (2008). Modeling cancer progression via pathway dependencies. *PLoS Comput. Biol.* 4(2), e28.
- Edelman, E., J. Guinney, A. Porello, B. Balakumaran, A. Bild, P. Febbo, and S. Mukherjee (2006). Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 22(14), e108–e116.
- 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.
- Fukumizu, K., F. Bach, and M. Jordan (2003). Kernel dimensionality reduction for supervised learning. In *Advances in Neural Information Processing Systems 16*.
- Fukumizu, K., F. Bach, and M. Jordan (2005). Dimensionality reduction in supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research* 5, 73–99.

- 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.
- 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*.
- Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon Press.
- Li, B., H. Zha, and F. Chiaromonte (2004). Linear contour learning: A method for supervised dimension reduction. pp. 346–356. UAI.
- 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. *statistica* 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*.
- Martin-Mérino, M. and J. Róman (2006). A new semi-supervised dimension reduction technique for textual data analysis. In *Intelligent Data Engineering and Automated Learning*.
- 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 (2009). Learning gradients and feature selection on manifolds.
- Mukherjee, S. and D. Zhou (2006). Learning coordinate covariances via gradients. *J. Mach. Learn. Res.* 7, 519–549.
- Nilsson, J., F. Sha, and M. Jordan (2007). Regression on manifolds using kernel dimension reduction. In *Proceedings of the 24th International Conference on Machine Learning*.
- Pillai, N., Q. Wu, F. Liang, S. Mukherjee, and R. Wolpert (2007). Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res.*, 1769–1797.
- 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.
- Speed, T. and H. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist. 14*, 138–150.
- Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *J. Mach. Learn. Res. 8*, 1027–1061.
- Tenenbaum, J., V. de Silva, and J. Langford (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*, 2319–2323.
- Tokdar, S., Y. Zhu, and J. Ghosh (2008). A bayesian implementation of sufficient dimension reduction in regression. Technical report, Purdue Univ.
- Tomlins, S., R. Mehra, D. Rhodes, X. Cao, L. Wang, S. Dhanasekaran, S. Kalyana-Sundaram, J. Wei, M. Rubin, R. Pienta, KJ and Shah, and A. Chinnaiyan (2007). Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics 39*(1), 41–51.
- Vlassis, N., Y. Motomura, and B. Kröse (2001). Supervised dimension reduction of intrinsically low-dimensional data. *Neural Computation*, 191–215.
- West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. In J. B. et al. (Ed.), *Bayesian Statistics 7*, pp. 723–732. Oxford.
- Wu, Q., J. Guinney, M. Maggioni, and S. Mukherjee (2007). Learning gradients: Predictive models that infer geometry and dependence. Technical Report 07, Duke University.
- 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.



(a) Y v.s. X_1, X_2, X_3

(b) X_3 v.s. X_1

Figure 1: Swiss Roll data: Illustration.

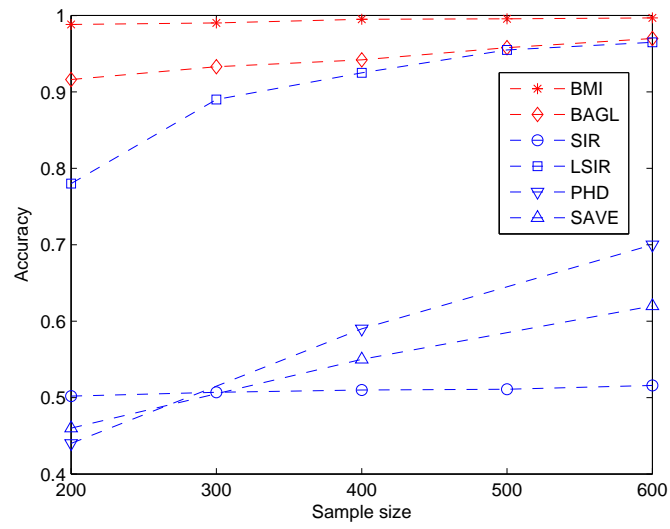


Figure 2: Swiss Roll data: Accuracy for different methods.

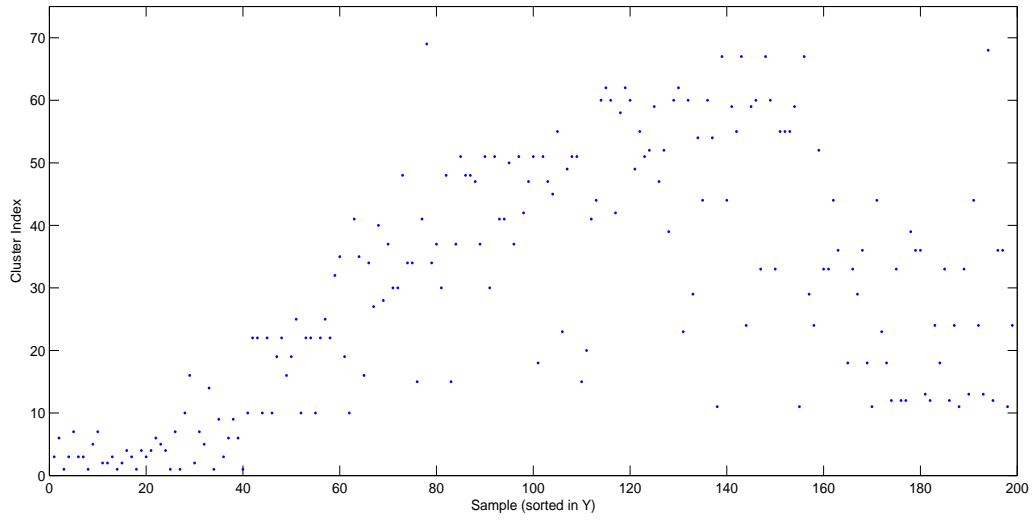
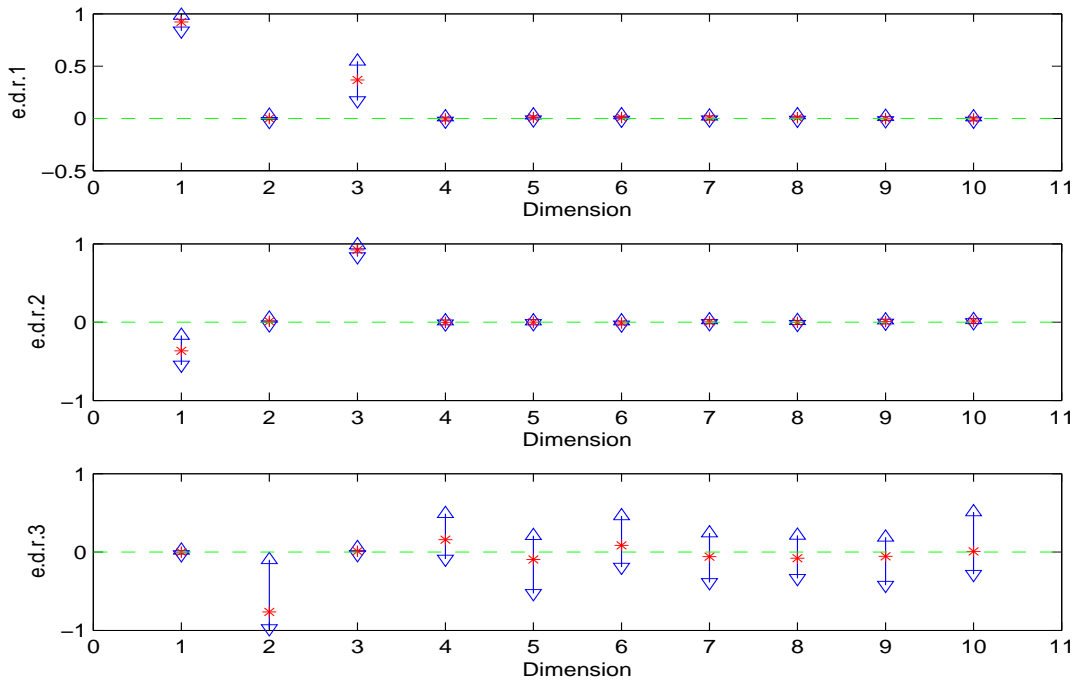
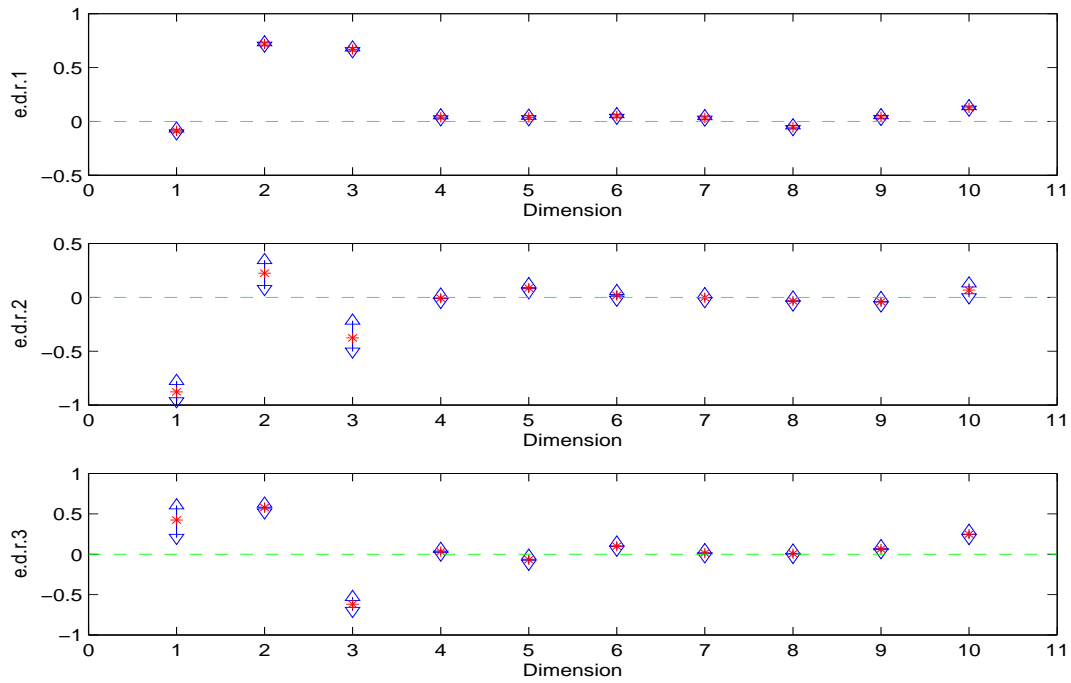


Figure 3: Cluster labels for all the 200 samples at the last iteration under an experiment. The samples are ordered in terms of the magnitude of Y . Closer samples (in terms of Y) seem to have similar underlying clustering distributions which change "gradually" with increasing response.

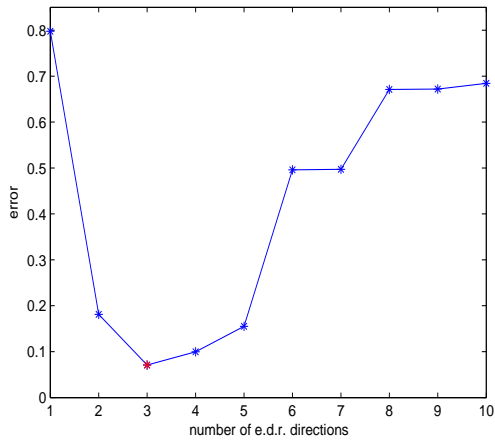


(a) BMI: Posterior mean and credible intervals

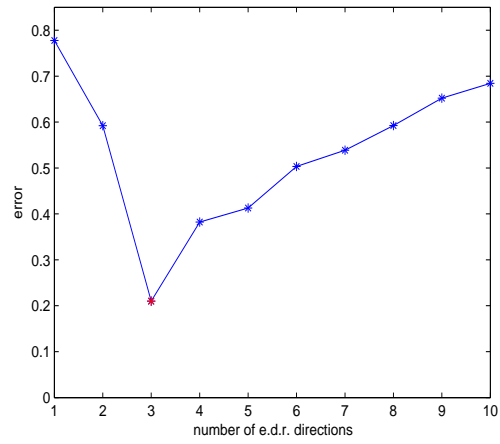


(b) BAGL: Posterior mean and credible intervals

Figure 4: Swiss Roll data: Posterior mean (red star) and 90% credible intervals (blue line segments) for the 3 e.d.r directions.

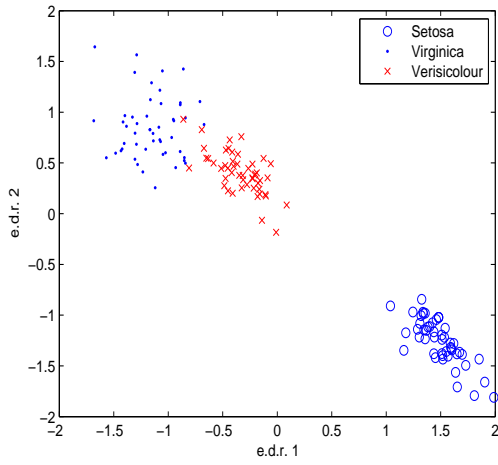


(a) BMI: Error v.s. d

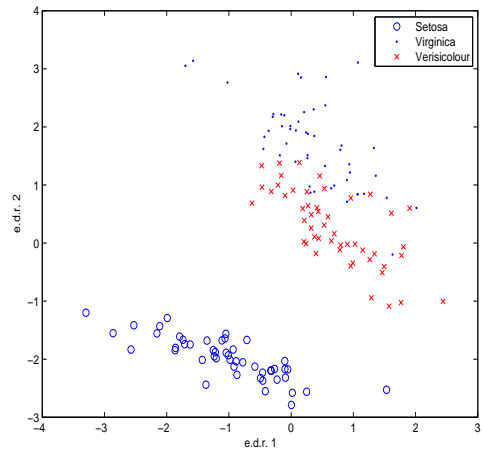


(b) BAGL: Error v.s. d

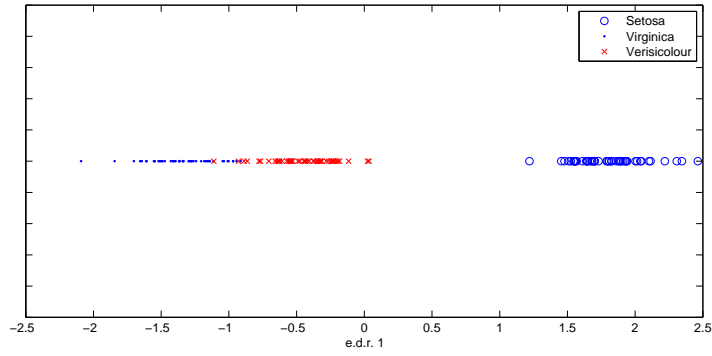
Figure 5: Swiss Roll data: Error v.s. number of e.d.r directions kept. The minimum one corresponds to $d = 3$, the true value.



(a) BMI: Embedded in 2 dimensional e.d.r. subspace



(b) BAGL: Embedded in 2 dimensional e.d.r. subspace



(c) BMI: Embedded in 1 dimensional e.d.r. subspace

Figure 6: Visualization of the embedded *Iris* data.

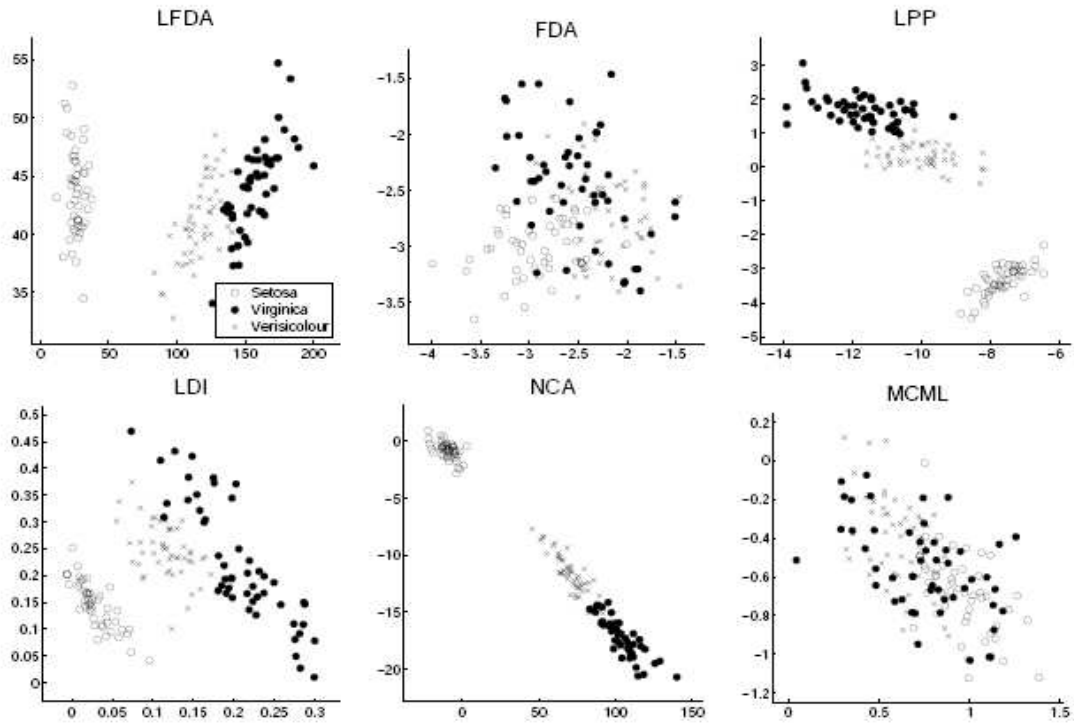
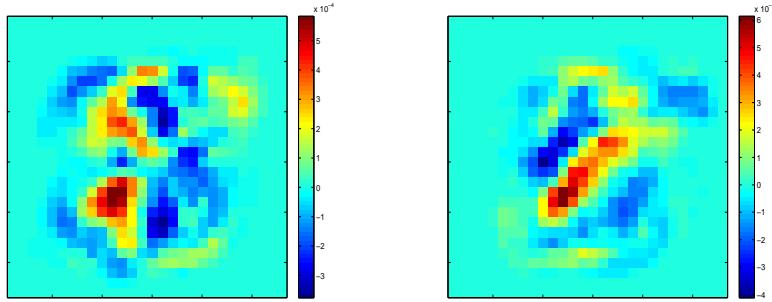


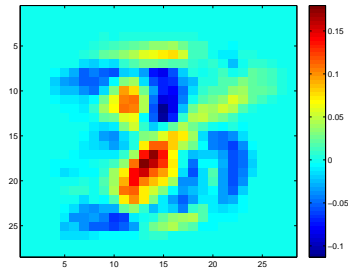
Figure 7: Visualization of the *Iris* data for different methods. (see also Sugiyama (2007) Figure 6.)



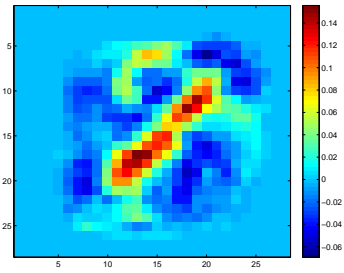
(a) Posterior mean of 3 vs 8

(b) Posterior mean of 5 vs 8

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

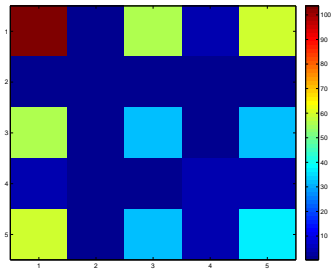


(a) Posterior mean of 3 vs 8

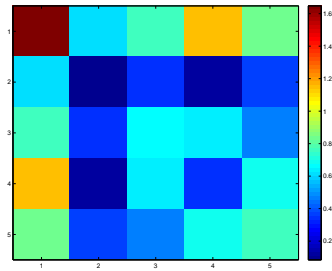


(b) Posterior mean of 5 vs 8

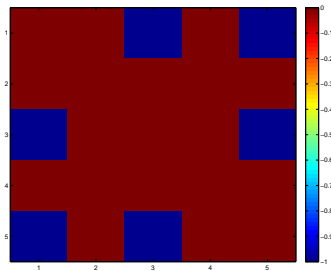
Figure 9: BAGL: (a) The posterior mean of the top feature for 3 versus 8, shown in a 28×28 pixel format. (b) The posterior mean of the top feature for 5 versus 8, shown in a 28×28 pixel format. Difference between digits is reflected by the red color.



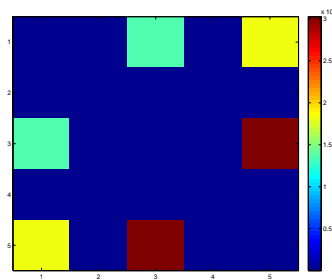
(a) Posterior mean of GOP



(b) Posterior standard deviation of GOP



(c) Posterior mean of partial correlation



(d) Posterior standard deviation

Figure 10: (a) and (b) are the posterior mean and standard deviation for the GOP, respectively; (c) and (d) are the posterior mean and standard deviation for the partial correlation matrix, respectively.

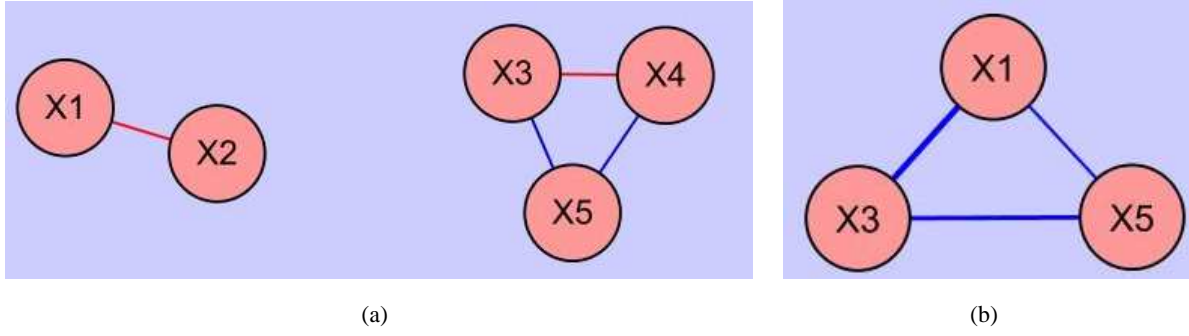


Figure 11: Graphical models inferred from the (a) the covariance matrix of the explanatory variables and (b) the gradient outer product matrix. Each node represents a variable and each edge indicates conditional dependence. The distance of the edge is inversely proportional to the amount of dependence, the thickness of the edge is proportional to the certainty of the inference and blue edges are negative while red edges are positive.

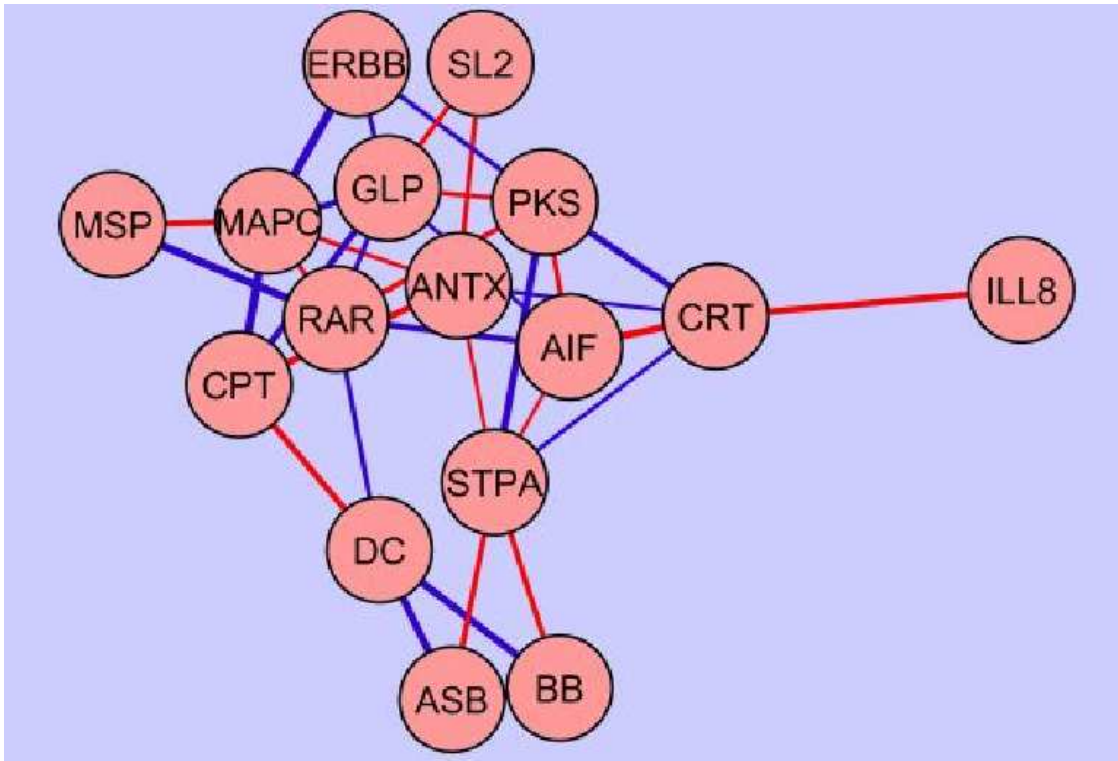


Figure 12: The association graph for the progression of prostate cancer from benign to malignant based on the inferred partial correlation. Red edges correspond to positive partial correlations and blue for negative. The width of the edges correspond to the degree of uncertainty, edges we are more sure of are thicker.