Sufficient Dimension Reduction via Bayesian Mixture Modeling

Brian J. Reich[1], Howard D. Bondell, and Lexin Li

Department of Statistics, North Carolina State University, Raleigh, NC

May 14, 2010

Summary.

Dimension reduction is central to an analysis of data with many predictors. Sufficient dimension reduction aims to identify the smallest possible number of linear combinations of the predictors, called the sufficient predictors, that retain all of the information in the predictors about the response distribution. In this paper we propose a Bayesian solution for sufficient dimension reduction. We directly model the response density in terms of the sufficient predictors using a finite mixture model. This approach is computationally efficient and offers a unified framework to handle categorical predictors, missing predictors, and Bayesian variable selection. We illustrate the method using both a simulation study and an analysis of an HIV data set.

Key words: Central subspace; Directional regression; Probit link function; Sliced inverse regression; Sufficient dimension reduction.

[1]Corresponding author, email: reich@stat.ncsu.edu.

1. Introduction

Dimension reduction is central to high-dimensional data analysis. For a regression of a response Y given a p-dimensional predictor vector X, sufficient dimension reduction (SDR, Cook, 1998) seeks a minimum subspace $\mathcal{S}$ of $\mathbb{R}^p$ whose basis $A \in \mathbb{R}^{p\times d}$ satisfies that $p(Y|X) = p(Y|A^\intercal X)$, where $p(\cdot)$ denotes the probability density function, and $d \leq p$ is the dimension of $\mathcal{S}$. In practice it is often the case that d is much smaller than p, so that substantial reduction is achieved. Reduction in this paradigm takes the form of the linear combinations $A^\intercal X$, which retain full regression information of $Y|X$ and are referred as sufficient predictors. Such a subspace is called the central subspace, is denoted by $\mathcal{S}_{Y|X}$, and offers a parsimonious characterization of the regression $Y|X$. Under very mild conditions (Cook, 1996, Yin, Cook and Li, 2008), the central subspace $\mathcal{S}_{Y|X}$ uniquely exists.

There have been many methods proposed to estimate $\mathcal{S}_{Y|X}$, which can be generally cast into two classes. One class relies on the inverse conditional moments of $X|Y$ to infer about $\mathcal{S}_{Y|X}$, and it requires some distributional assumptions on the predictors X, most notably, the elliptical symmetry condition. After obtaining a basis estimate $\hat{A}$ of $\mathcal{S}_{Y|X}$, the method passes the induced sufficient predictors $\hat{A}^\intercal X$ to subsequent model formulation and prediction. Examples of this class include the seminal sliced inverse regression (SIR, Li, 1991), sliced average variance estimation (Cook and Weisberg, 1991), and the more recent proposal of directional regression (Li and Wang, 2007), among others. The other class hinges on multi-dimensional kernel smoothing to simultaneously estimate the basis A and

the probability function $p(Y|A^\tau X)$. Examples include minimum average variance estimation (MAVE, Xia et al., 2002), a constructive estimator (Xia, 2007), and sliced regression (Wang and Xia, 2008). Nearly all existing SDR estimators take the frequentist approach, while Tokdar, Zhu and Ghosh (2008) recently considered a Bayesian implementation of SDR.

In this article we propose a Bayesian sufficient dimension reduction method by placing a prior on the central subspace and directly modeling the conditional distribution of the response in terms of the sufficient predictors. Similar to the previous approaches such as SIR and alike, the conditional density model partitions the response into homogeneous groups, or say, clusters. The cluster probabilities are related to the latent sufficient predictors via a probit link function. Then the probit link, combined with standard auxiliary variable techniques, relates the covariates to the conditional density via a standard multiple linear regression. We show later in Theorem 1 that this model actually spans a wide class of conditional densities in a limiting case.

There are several advantages of doing SDR this way. First, the group allocation is treated as random and depends on both the response and the predictors, as opposed to standard slicing which categorizes the observations based only on the response. In our solution the central subspace estimate is obtained by averaging over the cluster memberships according to their posterior probabilities. Model averaging is expected to provide more stable estimation of the central subspace. Second, similar to the class of MAVE estima-

tors, our proposal simultaneously estimates both the central subspace and the probability density of the response given the sufficient predictors. This one-step estimation is shown to be often more advantageous than the two-step procedure of obtaining a reduction estimate first then followed by a subsequent modeling (Xia, 2007). We have verified this advantage in our simulation study. On the other hand, our method avoids the use of a multi-dimensional kernel that is required by the class of MAVE estimators, and thus alleviates to some extent the curse of dimensionality. Third, it is well known that both classes of existing SDR methods would suffer if the predictors are categorical or a mixture of categorical and continuous variables. By contrast, our method does not impose any restriction on the distribution of X and can naturally deal with both types of predictors. We later use an AIDS study data to illustrate this, where there exist both continuous variables such as CD4 count and weight and binary variables such as gender and symptomatic status. Fourth, our Bayesian treatment of SDR makes the tackling of a variety of complications in real data straightforward. For instance, we can handle missing values in predictors, or incorporate domain knowledge of predictor structures, by employing appropriate priors. Finally, compared with the Bayesian SDR method of Tokdar, Zhu and Ghosh (2008) we model the conditional density using a mixture model rather than a logistic Gaussian process. Mixture models are often used for conditional density estimation (Jordan and Jacobs, 1994; Griffin and Steel, 2006; Dunson and Park, 2007; Chung and Dunson, 2010; among others). Our model is entirely conjugate which leads to simple coding and tuning. Also,

after introducing auxiliary variables, the model resembles standard multivariate linear regression, and therefore standard methods apply, for example, to account for missing data or perform variable selection.

The rest of the article proceeds as follows. We present our Bayesian dimension reduction model along with full conditionals in Section 2. We discuss identification and prior specification in Section 3, the central subspace estimation and variable selection in Section 4, and the MCMC algorithm in Section 5. We examine the empirical performance of the proposed method via simulations in Section 6 and a real data analysis in Section 7. We conclude the paper with a discussion in Section 8, including a summary of potential approaches for estimating the dimension of the central subspace, which we assume in known throughout the paper. Proofs and technical derivations are delegated to an Appendix.

2. Bayesian Dimension Reduction Model

2.1 *Single-index model*

Let y_i denote the response and $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ denote the p-dimensional predictors for observations $i = 1, \ldots, n$. Let A denote a basis of $\mathcal{S}_{Y|X}$ and $\lambda_i = (\lambda_{i1}, \ldots, \lambda_{id})^\intercal = A^\intercal \mathrm{x}_i$ denote the latent sufficient predictors. We begin describing our model assuming a single index, $d = 1$, so that λ_i is a scalar. To capture effects on higher moments, we specify a

flexible model for $p(y_i|\lambda_i)$, i.e., we model the conditional distribution as a finite mixture

$$p(y_i|\lambda_i) = \sum_{k=1}^{M} p_k(\lambda_i) N\left(\mu_k, \sigma_y^2\right), \tag{1}$$

where the mixture weights $p_k(\lambda_i)$ satisfy $\sum_{k=1}^{M} p_k(\lambda_i) = 1$ for all $\lambda_i \in \mathbb{R}^1$. For computational and conceptual convenience, we assume that the mixture probabilities can be modeled with the probit form

$$p_k(\lambda_i) = \Phi\left(\frac{\phi_{k+1} - \lambda_i}{\sigma_z}\right) - \Phi\left(\frac{\phi_k - \lambda_i}{\sigma_z}\right), \tag{2}$$

where Φ is the standard normal distribution function and $\phi_1 < \phi_2 < \ldots < \phi_{M+1}$ are cutpoints with $\phi_1 = -\infty$ and $\phi_{M+1} = \infty$. This permits the usual data augmentation. Introducing latent continuous variables z_i, the model can be written

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu_{g_i}, \sigma_y^2) \\ g_i = k \quad &\text{if} \quad \phi_k < z_i < \phi_{k+1} \\ z_i &\sim \mathrm{N}(\lambda_i, \sigma_z^2), \end{aligned} \tag{3}$$

which gives a fully-conjugate model and facilitates rapid MCMC sampling and convergence.

Figure 1 plots the conditional quantiles for one illustrative set of means μ_k and cut points ϕ_k. This figure shows how this relatively simple mixture model can capture complex features of the conditional distribution. For this example, the conditional mean is quadratic

and the variance is constant for a positive λ, whereas observations with a negative λ have mean is near zero and the variance is increasing as λ decreases.

To illustrate the generality of the conditional density model induced by the probit weights, we consider the case of an increasing grid of equally-spaced cutpoints $a = \psi_2 < \psi_3 < \ldots < \psi_M = b$ spanning the interval $[a, b]$ (recall $\psi_1 = -\infty$ and $\psi_{M+1} = \infty$). Let $\sigma_z = c\Delta_M$, where $c > 0$ and $\Delta_M = (b-a)/M$ is the grid width. Define $\psi_k^* = (\psi_k + \psi_{k+1})/2$ as the grid midpoint, $U_M(\lambda) = (\lambda - c\sqrt{\Delta_M}, \lambda + c\sqrt{\Delta_M})$ as a shrinking interval around λ, $\mathcal{S}(\lambda) = \{k|\psi_k^* \in U_M(\lambda)\}$ as the set of indices with midpoints in $U_M(\lambda)$, and $|\mathcal{S}(\lambda)|$ as the cardinality of $\mathcal{S}(\lambda)$.

Theorem 1 *As $M \to \infty$,*

(i) $\sum_{k\in\mathcal{S}(\lambda)} p_k(y|\lambda) \to 1$ and $|\mathcal{S}(\lambda)| \to \infty$.

(ii) For any $\lambda_1, \lambda_2 \in (a, b)$ so that $\lambda_1 \neq \lambda_2$, $U_M(\lambda_1) \cap U_M(\lambda_2) = \emptyset$.

The proof is given in the Appendix. As the number of cutpoints M increases and the probit variance σ_z^2 decreases at a particular rate, (i) states that the conditional density at λ is affected only by infinitely many terms with midpoints in the interval $U_M(\lambda)$, and (ii) states that the sets of terms affecting the conditional distributions at any two points are disjoint. Therefore, for a fixed sample size, this limiting case can fit a separate countable mixture of normals for the conditional density for each observation, and hence can be arbitrarily flexible. It is not our intention to fit or even approximate this limiting model, but this theorem illustrates that our model spans a wide class of conditional densities.

Using the probit model in (2), $p_k(\lambda)$, and thus the conditional density (1), are infinitely-differentiable functions of λ. Less smooth conditional densities could be modeled by replacing the Gaussian distribution function in (2) with a less smooth function. For example, a uniform distribution function would lead to a discontinuous conditional density, which may be desirable if the conditional density changes dramatically above a certain threshold for the latent variable λ. Since we believe the conditional density is fairly smooth for the data considered in the paper we use the probit weights, and we note that Theorem 1 shows that for small σ_z the conditional distribution model using probit weights is quite flexible.

The conditional density (1) models the response distribution as a function of the locations μ_k. To complete the model specification, these locations are considered as random with $\mu_k \overset{iid}{\sim} F_0$. In general it is difficult to examine the distributional properties of this prior for the conditional density. It is common in Bayesian non-parametrics to consider the limiting case with $\sigma_y = 0$. In this case, closed-form expressions are available for the mean and covariance of the CDF $F_\lambda(c) = P(y < c|\lambda)$ in terms of the base distribution $F_0(c) = P(\mu_k < c)$.

Theorem 2 *If $\sigma_y = 0$, the first two moments of $F_\lambda(c)$ are*

$$\begin{aligned} E(F_\lambda(c)) &= F_o(c) \\ Cov(F_{\lambda_1}(c), F_{\lambda_2}(c)) &= F_o(c)[1 - F_o(c)] \sum_{k=1}^{M} p_k(\lambda_1) p_k(\lambda_2). \end{aligned}$$

The prior mean shows that the conditional distribution is centered on the base distribution for all λ. The prior variance is a function of both the base distribution through $F_o(c)[1-F_o(c)]$ and the weights though $\sum_{k=1}^{M} p_k(\lambda)^2$. The variance is maximized at $F_o(c)[1-F_o(c)]$ when there is a single component with $p_k(\lambda) = 1$ and the remaining probabilities are zero. The variance goes to zero when there are many terms with small probability; in this case the base distribution plays an important role. For the remainder of the paper, we assume a normal base distribution F_o, so that $\mu_k \overset{iid}{\sim} N(0, \sigma_\mu^2)$.

To show how the the covariance decays as a function of $\lambda_1 - \lambda_2$, consider the limiting case with an equally-spaced grid of M points with grid spacing Δ_M. Then for large M

$$\mathrm{Cov}\left(F_{\lambda_1}(c), F_{\lambda_2}(c)\right) \approx \frac{F_o(c)[1-F_o(c)]}{2\sigma_z\sqrt{\pi}\Delta_M^2} \exp\left(-\frac{(\lambda_1-\lambda_2)^2}{4\sigma_z^2}\right).$$

Therefore, for a dense grid of points the covariance declines exponentially in $(\lambda_1 - \lambda_2)^2$.

2.2 *Multiple-index model*

To extend the single-index model to the multiple-index model with $d > 1$, we introduce a separate latent probit parameter for each dimension, z_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Let

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu_{g_{i1},\ldots,g_{id}}, \sigma_y^2) \\ g_{ij} = k \quad &\text{if} \quad \phi_k < z_{ij} < \phi_{k+1} \\ z_{ij} &\sim \mathrm{N}(\lambda_{ij}, \sigma_{zj}^2), \end{aligned} \tag{4}$$

where $\mu_{k_1,\ldots,k_d} \overset{iid}{\sim} \mathrm{N}(0, \sigma_\mu^2)$. In this notation, $\mu_{k_1,\ldots,k_d}$ is a scaler with d indices, one for each dimension of the central subspace. Integrating over the latent variables z_{ij} gives

$$p_{k_1,\ldots,k_d}(\lambda_i) = \prod_{j=1}^{d} \left[\Phi\left(\frac{\phi_{k_j+1} - \lambda_{ij}}{\sigma_{zj}} \right) - \Phi\left(\frac{\phi_{k_j} - \lambda_{ij}}{\sigma_{zj}} \right) \right], \tag{5}$$

and thus

$$p(y_i|\boldsymbol{\lambda}_i) = \sum_{k_1=1}^{M} \cdots \sum_{k_d=1}^{M} p_{k_1,\ldots,k_d}(\boldsymbol{\lambda}_i) N\left(\mu_{k_1,\ldots,k_d}, \sigma_y^2\right). \tag{6}$$

As discussed in Section 5, this model is also fully-conjugate.

It is straight-forward to extend Theorem 1 to this multiple index model, illustrating the flexibility of the model in higher dimensions. However, this model is mixture of M^d components, which may be inefficient for even moderate d. One approach to reducing the dimension is an ANOVA representation of the mixture means

$$\mu_{k_1,\ldots,k_d} = \sum_{j=1}^{d} \alpha_{k_j}^{(j)} + \sum_{l<j} \gamma_{k_l,k_j}^{(l,j)} + \ldots \tag{7}$$

where $\alpha^{(j)} \sim \mathrm{N}(0, \sigma_{\alpha j}^2)$ are the main effects for index j, $\gamma^{(l,j)} \sim \mathrm{N}(0, \sigma_{\gamma lj}^2)$ are the two-way interactions effects for indices l and j, and so on. Including up to d-way interactions gives the model with the same dimension as the original model. Using only lower-dimensional interactions reduces the dimension of the model, for example retaining only two-way interactions reduces the number of terms from M^d to $M[d + \binom{d}{2}]$.

Consider the model without interactions, $\mu_{k_1,\ldots,k_d} = \sum_{j=1}^{d} \alpha_{k_j}^{(j)}$. The mean response conditioned on λ and $\alpha_k^{(j)}$ becomes

$$E(y_i|\boldsymbol{\lambda}_i, \alpha_k^{(j)}) = \sum_{k=1}^{M} p_k(\lambda_{i1}|\sigma_{z1})\alpha_k^{(1)} + \ldots + \sum_{k=1}^{M} p_k(\lambda_{id}|\sigma_{zd})\alpha_k^{(d)}, \tag{8}$$

where $p_k(\lambda_{ij}|\sigma_{zj}) = \Phi\left(\frac{\phi_{k+1}-\lambda_{ij}}{\sigma_{zj}}\right) - \Phi\left(\frac{\phi_k-\lambda_{ij}}{\sigma_{zj}}\right)$. Therefore, the conditional mean resembles the generalized additive model (Hastie and Tibshirani, 1990) with $p_k(\lambda_{ij}|\sigma_{zj})$ and $\alpha_k^{(j)}$ playing the roles of the basis functions and basis function coefficients, respectively, for dimension j. Replacing the sigmoid function with the probit function, the form of the basis functions $p_k(\lambda_{ij}|\sigma_{zj})$ resemble that of neural networks (Hastie, Tibshirani, and Friedman, 2001), in that the mean is a linear combination of non-linear functions of linear combinations of the predictors.

In general, there is no guarantee that the central subspace can be recovered if a lower-dimensional model is assumed and the true distribution cannot be well-approximated by a lower-dimensional model. In this case, there is a trade-off between parsimony and flexibility. That is, if the dimension of the central subspace is large and the relationship between the sufficient predictors and the response is complicated, a very large and flexible model is needed to fit the data. However, this large model is likely to be poorly-identified unless the sample size is very large. Thus it is not clear that a flexible model is preferred even when the true model is complicated.

3. Model Identification and Prior Specification

Clearly there is non-identifiability between the cutpoints ϕ and the latent effects λ, since adding a constant to all the cutpoints and latent effects would give the same probabilities p_k. Therefore the priors for ϕ and A must be chosen carefully to ensure MCMC convergence. To address this non-identifiability, we assume the predictors are centered and scaled to have mean zero and variance one. Let $A = \{a_{lj}\}$, where a_{lj} is the effect of the l-th covariate on the j-th latent effect λ_{ij}. The scale of A does not affect the central subspace, so we assume a_{lj} is normal with mean zero and variance $\sigma^2_{alj} \leq 1$, independent across l and j. Then for large p and independent predictors the prior of λ_{ij} is approximately normal with mean zero and variance less than or equal to p. The cutpoints can either be fixed or modeled as random with an appropriate prior. The cutpoints must be chosen or modeled to span the range of $\lambda_{ij} \sim \mathrm{N}(0, p)$, roughly the interval $(-3\sqrt{p}, 3\sqrt{p})$. Therefore, in all analyses we assume that the cutpoints ϕ are fixed on a grid of M equally-spaced quantiles of the $N(0, 9p)$ distribution.

For the multiple-index model, we assume that A is lower triangular. This improves MCMC convergence while retaining the full class of central spaces in the prior. Ideally the columns with elements set to zero would be those that are most representative of the previous directions. Therefore we choose the variables to have columns set to zero sequentially. We first fit the single-index model and identify the variable with largest posterior mean a^2_{l1}, and set $a_{l2} = \ldots = a_{ld} = 0$ for that covariate. Then we fit the model with

$d = 2$ and identify the variable with largest posterior mean a_{l2}^2, and set $a_{l3} = \ldots = a_{ld} = 0$ for that covariate, and so on. Prior for the non-zero elements of A are discussed in Section 4.2

We use a normal distribution with mean μ_0 and variance σ_μ^2 for the base distribution F_0. The overall mean μ_0 has a $N(0, \sigma_0^2)$ prior. The variances σ_y^2, σ_μ^2 (or $\sigma_{\alpha j}^2$ and $\sigma_{\gamma lj}^2$), and σ_{zj}^2 have independent InvGamma(ϵ, ϵ) priors, parameterized to have mean 1 and variance $1/\epsilon$. We take $\sigma_0 = 10^4$ and $\epsilon = 0.1$. In Section 7 we conduct sensitivity analyses for these assumptions and find that for moderate sample sizes the results are not sensitive to these priors.

Finally, we discuss selecting the number of terms, M. Selecting M to be too small disregards information in the data, and selecting M to be too large gives an over-parameterized model. One option would be to treat M as an unknown quantity and use reversible jump MCMC to account for uncertainty in M. Alternatively, M could selected using cross-validation or some other model-selection criterion. However, as in Wang and Xia (2008), we find that for a reasonable range, say M between 5 and 10, the results are not sensitive to M. We therefore opt for the simplest approach of fixing M and conducting a sensitivity analysis of the effect of M on the estimate of the central subspace, as demonstrated in Section 7.

4. Dimension Reduction Estimation and Variable Selection

4.1 *Estimation of the central subspace*

The objective of dimension reduction is to estimate the central subspace, i.e., the span of A. The central subspace is difficult to quantify using draws from the posterior as it is invariant to the A's sign, scale, and column order. Therefore, summarizing the posterior using the component-wise posterior mean or median of A is invalid because the sign of A may change from sample to sample, so the posterior mean of an element of A may be zero even if there are no draws near zero. To resolve these issues, we propose to estimate the central subspace using the span of the first d eigenvectors of the component-wise posterior mean of the projection matrix $P = A(A'A)^{-1}A$. Below we show that this is the Bayes estimator with respect to the Frobenius norm loss.

Let F be a distribution over $\mathcal{P}_d$, the space of all $p \times p$ orthogonal projection matrices of rank d, and let $P \in \mathcal{P}_d$ denote a random draw from F. Denote the expectation with respect to F by $P_\mu = EP$, and let $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_p$ be the ordered eigenvalues of P_μ, and U the matrix whose columns are the corresponding eigenvectors. Note that $P_\mu \notin \mathcal{P}_d$, in general. The proof of the next theorem is given in the Appendix.

Theorem 3 *Define* $P_0 = \arg\min_{Q \in \mathcal{P}_d} E||P - Q||^2$*, where* $||\cdot||$ *denotes the Frobenius norm. Then* $P_0 = U_d U_d'$*, where* U_d *denotes the first* d *columns of* U*, i.e., the projection onto the span of the first* d *eigenvectors of* $P_\mu = EP$*.*

Corollary 1 *The Bayes estimator with respect to the loss function* $||P - Q||^2$ *is* $Q = U_d U_d'$*,*

where U_d represents the eigenvectors corresponding to the first d eigenvalues of the posterior mean of P.

4.2 *Variable selection*

When the number of predictors is large, estimates of the central subspace may be more stable after eliminating unnecessary predictors. We consider two choices for the prior variance of the non-zero elements of A. The first fixes $\sigma^2_{alj} = 1$ for all l and j. In cases with a large number of predictors, it may be desirable to encourage sparsity. To eliminate null predictors, we also consider stochastic search variable selection (SSVS) via the two component mixture prior (George and McCulloch, 1993) $\sigma^2_{alj} = \pi_{lj} + c^2(1 - \pi_{lj})$, where $\pi_{lj} \sim \text{Bern}(\bar{\pi})$ and $0 < c < 1$ is a small fixed constant.

The motivation for the two-component mixture model is as follows. The binary parameter π_{lj} represents whether the l-th covariate is active for dimension j; if π_{lj} is one the l-th covariate is included in the model for dimension j, and if $\pi_{lj} = 0$ then a_{lj}'s prior variance is c^2 and the l-th covariate is effectively removed from the model for dimension j. Rather than select a particular π_{lj} and present the results of a single model, in the SSVS approach we treat the model as an unknown quantity updated in the MCMC algorithm. By doing so, we obtain an estimate of the posterior probability of each candidate model indexed by π_{lj}, and our estimate of the central subspace is averaged over models according to their posterior probability. We declare variable l as being included in the model if the posterior probability

that it is included in at least one dimension, i.e., P($\max\{\pi_{l1}, ..., \pi_{ld}\} = 1$|data), is greater than 0.5. To ensure covariates with $\pi_{lj} = 0$ have coefficients near zero we take $c = 0.01$, and to give a vague prior on the inclusion probabilities we assume $\bar{\pi} \sim$ Uniform(0,1).

5. Computations

The full conditionals are described below for the single-index model. For notational convenience, define $\tau_y = 1/\sigma_y^2$, $\tau_\mu = 1/\sigma_\mu^2$, $\tau_z = 1/\sigma_z^2$, $\tau_0 = 1/\sigma_0^2$, $\tau_{\gamma k} = 1/\gamma_k^2$ and let $g_i \in \{1, \ldots, M\}$ indicate the mixture component of the i-th observation, so that $g_i = k$ if $\phi_k < z_i < \phi_{k+1}$. The full conditionals for the parameters in the likelihood are

$$\begin{aligned}
\mu_k|\text{rest} &\sim \text{Normal}\left(\frac{\tau_{\gamma k}\mu_0 + \tau_y \sum_{i=1}^n I(g_i = k) y_i^2}{\tau_{\gamma k} + \tau_y \sum_{i=1}^n I(g_i = k)}, \frac{1}{\tau_{\gamma k} + \tau_y \sum_{i=1}^n I(g_i = k)}\right) \\
\tau_y|\text{rest} &\sim \text{Gamma}\left(n/2 + \alpha, \sum_{i=1}^n (y_i - \mu_{g_i})^2/2 + \beta\right).
\end{aligned}$$

The hyperparameters for μ_k have full conditionals

$$\begin{aligned}
\tau_{\gamma k}|\text{rest} &\sim \text{Gamma}\left(1/2 + \nu/2, (\mu_k - \mu_0)^2/2 + \nu\tau_\mu/2\right) \\
\mu_0|\text{rest} &\sim \text{Normal}\left(\frac{\sum_{k=1}^M \tau_{\gamma k}\mu_k^2}{\sum_{k=1}^M \tau_{\gamma k} + \tau_0}, \frac{1}{\sum_{k=1}^M \tau_{\gamma k} + \tau_0}\right) \\
\tau_\mu|\text{rest} &\sim \text{Gamma}\left(\nu/2 + \epsilon, \sum_{k=1}^M \tau_{\gamma k}\nu/2 + \epsilon\right), \\
P(\nu = k)|\text{rest} &\propto \prod_{k=1}^M G\left(\tau_{\gamma k}|k/2, k\tau_\mu/2\right),
\end{aligned}$$

where $G(y|a,b)$ is the gamma density function. The hyperparameters for the priors of the latent $\mathrm{z} = (z_1, \ldots, z_n)^\intercal$ have full conditionals

$$\begin{aligned} A|\text{rest} &\sim \text{Normal}\left(\tau_z(\tau_z \mathrm{X}'\mathrm{X} + S_a)^{-1}\mathrm{X}'\mathrm{z}, (\tau_z \mathrm{X}'\mathrm{X} + S_a)^{-1}\right) \\ \tau_z|\text{rest} &\sim \text{Gamma}\left(n/2 + \alpha, \sum_{i=1}^{n}(z_i - \lambda_i)^2/2 + \beta\right), \end{aligned}$$

where S_a is diagonal with diagonal elements σ_{ak}^{-2}, $\mathrm{z} = (z_1, ..., z_n)'$ and $\mathrm{X} = (\mathrm{x}_1', ..., \mathrm{x}_n')'$.

The inclusion indicators have full conditional

$$\pi_l|\text{rest} \sim \text{Bernoulli}\left(\frac{\bar{\pi}N(a_l|0,1)}{\bar{\pi}N(a_l|0,1) + (1-\bar{\pi})N(a_l|0,c^2)}\right),$$

where $N(a|m,v)$ is the Gaussian density with variate a, mean m, and variance v.

Finally, the latent z_i's full conditional is a mixture of truncated normals. We sample from this distribution by first drawing the mixture component g_i with

$$P(g_i = k|\text{rest}) \propto p_k(\lambda_i)\exp\left(-\tau_y(y_i - \mu_k)^2/2\right),$$

and then drawing z_i from a normal distribution with mean λ_i and variance σ_z^2, truncated to $(\phi_{g_i}, \phi_{g_i+1})$. The full conditionals for the multiple index model are nearly identical, expect that z_i's full conditional is a mixture of M^d truncated normals. The details are thus omitted here.

MCMC sampling is carried out using R, while it would also be straightforward to implement the model using WinBUGS. We draw 20,000 MCMC samples and discard the first 5,000 as burn-in. Convergence is monitored using trace plots of the deviance as well as several representative parameters. Since the model is entirely conjugate, we use Gibbs sampling to update all parameters. Code is available from the first author by request.

6. Simulation Study

In this section we conduct a simulation study to compare our method with the long-standing sliced inverse regression (SIR, Li, 1991) and the state-of-the-art directional regression (DR, Li and Wang, 2007) and sliced regression (SR, Wang and Xia, 2008). We replicate the simulation designs of Li and Wang (2007).

$$\begin{aligned}
&\text{Design 1:} && Y = 0.4V_1^2 + 3\sin(V_2/4) + \varepsilon,\\
&\text{Design 2:} && Y = 3\sin(V_1/4) + 3\sin(V_2/4) + \varepsilon,\\
&\text{Design 3:} && Y = 0.4V_1^2 + \sqrt{|V_2|} + \varepsilon,\\
&\text{Design 4:} && Y = 3\sin(V_2/4) + [1 + V_1^2]\varepsilon.
\end{aligned}$$

The predictors X are generated independently from the standard normal distribution. The two linear predictors $V_1 = A_1^\intercal X$ and $V_2 = A_2^\intercal X$, where $A_1 = (1, 1, 1, 0, \ldots, 0)^\intercal$ and $A_2 = (1, 0, 0, 0, 1, 3, 0, \ldots, 0)^\intercal$. The error ε is normal with mean 0 and standard deviation

0.2. We generate 100 data sets from each model with $n = 100$ and $p = 6$, and with $n = 500$ and $p = 20$. We also consider a non-additive model from Li (1991) with $n = 400$, $p = 10$, $A_1 = (1, 0, \ldots, 0)^\intercal$, $A_2 = (0, 1, 0, \ldots, 0)^\intercal$, and

$$\text{Design 5:} \qquad Y = \frac{V_1}{0.5 + (V_2 + 1.5)^2} + \varepsilon.$$

In addition to the frequentist solutions of SIR, DR, and SR we also examine three variations of our Bayesian dimension reduction (BDR) methods: (a) the proposed BDR with both ANOVA decomposition and variable selection, where $\bar{\pi} \sim$ Uniform(0,1) and $\alpha_k^{(j)} \neq 0$ for all j and k; (b) the BDR method with no variable selection (BDR-I), where $\bar{\pi} = 1$ and $\alpha_k^{(j)} \neq 0$ for all j and k; and (c) the BDR method without ANOVA decomposition (BDR-II), where $\bar{\pi} \sim$ Uniform(0,1) and $\alpha_k^{(j)} \equiv 0$ for all j and k. For models BDR and BDR-I that use the ANOVA decomposition, we include both main effects and two-way interactions with priors described below (7). We choose $M = 7$ cutpoints for the BDR methods. Following the implementation as in Wang and Xia (2008), the final bandwidth is given by $h_d = 0.1\kappa n^{-1/(d+4)}$, where d is the dimension of the reduction subspace. Wang and Xia, used two choices, $\kappa = 5$ and $\kappa = 10$. In addition, they considered two choices of H, the number of slices, $H = 5$ and $H = 10$. For our simulations, we used each of the four combinations for the pair (H, κ) and reported the best of the four results for each example. For DR and SR we fit both 5 and 10 slices and the report results for the number of slices

with the best performance for each design.

To evaluate the estimation accuracy, we compute the matrix distance between the projections to the true and the estimated central subspaces, (Li, Zha, and Chiaromonte, 2005): trace$\{(\mathrm{P}-\hat{\mathrm{P}})(\mathrm{P}-\hat{\mathrm{P}})\}$, where $\mathrm{P}=A(A^{\intercal}A)^{-1}A$, $\mathrm{P}=\hat{A}(\hat{A}^{\intercal}\hat{A})^{-1}\hat{A}$, and A and $\hat{A}$ denote the true and estimated basis matrix of $\mathcal{S}_{Y|X}$. We report the mean and standard error (in parenthesis) of this matrix distance over 100 data replications in Table 1.

We first see from the table that our Bayesian solution with both variable selection and the ANOVA decomposition ("BDR") achieves a better estimation accuracy compared to SIR and DR in all designs. When the conditional mean $E(Y|X)$ is additive in terms of the latent sufficient predictors (Designs 1-3), the ANOVA decomposition of the mixture means improves estimation of the central subspace. For Design 2 with $n = 100$, the Bayesian methods without ANOVA decomposition has larger mean matrix distance than the frequentist methods. For the range of data in this study, $\sin(V_j/4) \approx V_j/4$, so without further assumptions about form of the model it is difficult separate the two directions for small sample sizes. When the true basis matrix is sparse ($p = 20$), BDR incorporating variable selection also improves the estimation accuracy. Table 2 reports the results in terms of variable selection, and it is clearly seen that our method performs well in eliminating irrelevant predictors.

7. Analysis of HIV data

We complement our simulation study by analyzing a real data from the AIDS Clinical Trials Group Protocol 175 (ACTG 175), where there are both continuous and categorical predictors. This data has been previously analyzed by Hammer et al. (1996), Leon et al. (2003), Davidian et al. (2005), and Tsiatis et al. (2008). The HIV-infected subjects are randomized to four different antiretroviral regimens in equal proportions: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine, and ddI monotherapy. We follow the previous analyses and consider two groups: ZDV monotherapy (532 subjects), and the other three groups combined (1607 subjects). The response is the CD4 count (cells/mm3) at 20 ± 5 weeks post-baseline. The $p = 12$ predictors are given in Table 3.

Our objective is to build a model for a subject's response under each treatment, and use the model to estimate the probability that the subject will have a better response under ZDV monotherapy than the other treatment. We apply our Bayesian dimension reduction method separately for each treatment group, where for the ease of model interpretation, we employ the single-index model version of our method. We use $M = 10$ mixture components, and the same priors as those given in Section 5. Table 3 gives the inclusion probabilities for the predictors for each treatment. For both treatments CD4 cell count is the strongest predictor and is included with probability one. Antiretroviral history also has inclusion probability one for each treatment. CD8 cell count and non-white race also have inclusion probabilities greater than 0.5 for at least one treatment.

We refit with only those predictors with inclusion probabilities greater than 0.5 for

at least one treatment. For prediction, we use the posterior mean of the A. Since the coefficient for CD4 cell counts is always positive, we estimate the central subspace directly using the sample mean of the MCMC draws of A. The estimated latent sufficient predictors for ZDV monotherapy (λ_0) and the other treatments (λ_1) are

$$\begin{aligned}
\lambda_0 &= 1.49 * \text{CD4} - 0.12 * \text{CD8} - 0.38 * \text{I(Antiretroviral history)} + 0.00 * \text{I(non-white)} \\
\lambda_1 &= 1.42 * \text{CD4} - 0.17 * \text{CD8} - 0.33 * \text{I(Antiretroviral history)} - 0.14 * \text{I(non-white)}
\end{aligned}$$

where the predictors are standardized using the sample mean and standard deviation given in Table 3. Note that λ_0 and λ_1 are not perpendicular because they are estimates of the single index for two different data sets, rather than estimates of two dimensions for the same subspace.

Figures 2a and 2b show the posterior mean conditional density as a function of the single latent variable. For both treatments, the conditional mean and variance both increase as a function of the latent variable. The conditional densities appear nearly Gaussian for small values of λ, and are non-Gaussian with heavy tails and skewness, especially for the treatments other than ZDV monotherapy, when λ is large.

Figure 2c plots the probability that the response is smaller for ZDV monotherapy than the other treatments and the estimated difference in mean response between treatments, as a function of the latent indices. This plot spans the range of values of λ_0 and λ_1 evaluated

for each subject in the data set. Since the latent variables are highly correlated for these subjects, the axes are rotated to be uncorrelated for the observed subjects. For most subjects the probability that the response is smaller for ZDV monotherapy than the other treatments is over 0.5. Over the range of covariates observed in these data this probability ranges from 0.56 to 0.70. The subjects with the highest probability are those with large $|\lambda_0 + \lambda_1|$, and with $\lambda_0 < \lambda_1$ (lower corners of Figure 2c). Inspection of the data reveals that these subjects are generally non-whites with low CD4 and CD8 cell counts. Figure 2d shows the estimated difference in the mean response between the treatment groups is between 30 and 120 cells/mm3; this generally agrees with previous analyses, e.g., Tsiatis et al. (2008).

Finally we conduct a brief analysis of prior sensitivity. We use only the ZDV monotherapy and the model with all $p = 12$ predictors. The analysis above used $M = 10$ mixture components and $\epsilon = 0.1$ in the inverse gamma prior for the variances. We inspect the change in inclusion probabilities and central subspace estimate due to changing M to multiples of 5 from $M = 5$ to $M = 50$, and changing ϵ to 0.01 and 1. The results are given in Figure 3. For this data, the estimates are not sensitive over this range of ϵ. There is some change in the estimates as M increases, but the results are qualitatively similar for $M > 5$. For example, for all M the same two variables (baseline CD4 and Antiretroviral history) clearly stand out of the most significant, and the variables included in the model with probability greater than 0.5 are the same for $M > 5$.

8. Discussion

In this paper we propose a Bayesian model for sufficient dimension reduction. Our approach is computationally convenient and naturally accommodates missing and categorical predictors and Bayesian variable selection. We show via a simulation study that the model has good small-sample performance compared to standard approaches.

We have assumed throughout the paper that the dimension of the central subspace, d, is known. Similar to the number of mixture components, there are other options. For example, one possibility is to allow the dimension to be an unknown quantity and use reversible jump MCMC to compute its posterior distribution. However, when implementing this approach, we encountered poor convergence. An alternative would be to select the dimension using a data-based criterion. For example, it can be possible to implement a cross-validation approach based on the mixture model in order to select the dimension. This is an important area of future work.

Appendix

Proof of Theorem 1. Statement (ii) holds since $U_M(\lambda)$ is constructed to be a shrinking towards $\{\lambda\}$ and $\lambda_1 \neq \lambda_2$. The statement (i) is satisfied since the mass in the interval U_M is $\sum_{k\in\mathcal{S}(\lambda)}^{M} p_k(y|\lambda) > \Phi(c\sqrt{\Delta_M}/\sigma_z) - \Phi(-c\sqrt{\Delta_M}/\sigma_z) = \Phi(1/\sqrt{\Delta_M}) - \Phi(-1/\sqrt{\Delta_M}) \to 1$ as $\Delta_M \to 0$, and the number of terms with midpoints in $U_M(\lambda)$ is $|\mathcal{S}(\lambda)| \to 2cM\sqrt{\Delta_M} = 2c\sqrt{M(b-a)}$.

Proof of Theorem 2. The prior mean is

$$E\left(F_\lambda(c)\right) = \sum_{k=1}^{M} p_k(\lambda) E(I(\mu_k < c)) = \sum_{k=1}^{M} p_k(\lambda) F_o(c) = F_o(c).$$

Also,

$$\begin{aligned} E\left(F_{\lambda_1}(c)F_{\lambda_2}(c)\right) &= \sum_{k=1}^{M}\sum_{l=1}^{M} p_k(\lambda_1)p_l(\lambda_2)F_o(c)^{I(k=l)} \\ &= \sum_{k=1}^{M}\sum_{l=1}^{M} p_k(\lambda_1)p_l(\lambda_2)F_o(c)^2 + \sum_{k=1}^{M} p_k(\lambda_1)p_l(\lambda_2)\left[F_o(c) - F_o(c)^2\right] \\ &= F_o(c)^2 + F_o(c)\left(1 - F_o(c)\right)\sum_{k=1}^{M} p_k(\lambda_1)p_l(\lambda_2). \end{aligned}$$

Therefore,

$$\mathrm{Cov}\left(F_{\lambda_1}(c), F_{\lambda_2}(c)\right) = E\left(F_{\lambda_1}(c)F_{\lambda_2}(c)\right) - E\left(F_{\lambda_1}(c)\right)E\left(F_{\lambda_2}(c)\right) = F_o(c)\left(1 - F_o(c)\right)\sum_{k=1}^{M} p_k(\lambda_1)p_l(\lambda_2).$$

In the limiting case of an equally-spaced grid of M points with spacing Δ_M and large M,

$$\begin{aligned} \sum_{k=1}^{M} p_k(\lambda_1)p_l(\lambda_2) &= \sum_{k=1}^{M}\left[\Phi\left(\frac{\phi_{k+1} - \lambda_1}{\sigma_z}\right) - \Phi\left(\frac{\phi_k - \lambda_1}{\sigma_z}\right)\right]\left[\Phi\left(\frac{\phi_{k+1} - \lambda_2}{\sigma_z}\right) - \Phi\left(\frac{\phi_k - \lambda_2}{\sigma_z}\right)\right] \\ &\approx \frac{1}{\Delta_M^2}\int N(\phi^*|\lambda_1, \sigma_z^2)N(\phi^*|\lambda_2, \sigma_z^2)d\phi^* \\ &= \frac{1}{2\sigma_z\sqrt{\pi}\Delta_M^2}\exp\left(-\frac{(\lambda_1 - \lambda_2)^2}{4\sigma_z^2}\right) \end{aligned}$$

Proof of Theorem 3. We first note that $E||P - Q||^2 = E||P - P_\mu||^2 + ||P_\mu - Q||^2$.

Hence $P_0 = \arg\min_{Q \in \mathcal{P}_d} ||P_\mu - Q||^2$. Now, $||P_\mu - Q||^2 = \text{trace}\{(P_\mu - Q)^\intercal (P_\mu - Q)\} = \text{trace}\{(P_\mu - Q)(P_\mu - Q)\}$, since both P_μ and Q are symmetric. Since Q is an orthogonal projection matrix of rank d, it follows that $\text{trace}(Q^2) = \text{trace}(Q) = d$. Hence $||P_\mu - Q||^2 = d + \text{trace}\left(P_\mu^2\right) - 2\text{trace}\,(QP_\mu)$. Thus, $P_0 = \arg\max_{Q \in \mathcal{P}_d} \text{trace}\,(QP_\mu)$. Using von Neumann's trace inequality (von Neumann, 1937), $\text{trace}\,(QP_\mu) \leq \sum_{j=1}^{p} q_j \lambda_j$, where $q_1 \geq q_2 \geq \cdots \geq q_p$ are the eigenvalues of Q. Since Q is an orthogonal projection matrix of rank d, it must be that $q_1 = \cdots = q_d = 1$ and $q_{d+1} = \cdots = q_p = 0$. Hence $\text{trace}\,(QP_\mu) \leq \sum_{j=1}^{d} \lambda_j$, but equality is obtained if $Q = U_d U_d^\intercal$.

References

Chung, Y. and Dunson, D.B. (2010). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, to appear.

Cook, R.D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983-992.

Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, **86**, 328-332.

Davidian M., Tsiatis A.A., Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with Discussion). *Statistical Science*, **20**, 261-301.

Dunson, D.B., Park, J-H. (2008). Kernel stick breaking processes. *Biometrika*, **95**, 307-323.

George, E.I., McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881-889.

Griffin, J. E., Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *ournal of the American Statistical Association*, **101**, 179-94.

Hammer, S.M., Katzenstein, D.A., Hughes, M.D., Gundaker, H., Schooley, R.T., Haubrich, R.H., Henry, W.K., Lederman, M.M., Phair, J.P., Niu, M., Hirsch, M.S. and Merigan, T.C., for the AIDS Clinical Trials Group Study 175 Study Team (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, **335**, 1081-1089.

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hastie, T.J., Tibshirani, R.J, and Friedman J. (2001). *The elements of statistical learning*. New York: Springer.

Jordan, M. I., Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181-214.

Leon, S., Tsiatis, A.A., Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, **59**, 1048-1057.

Li, B., and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997-1008.

Li, B., Zha, H., and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, **33**, 1580-1616.

Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-327.

Tokdar, S.T., Zhu, T, Ghosh, J.K. (2008). A Bayesian Implementation of Sufficient Dimension Reduction in Regression. Technical report, Carnegie Mellon University.

Tsiatis, A.A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine*, **27**, 4658-4677.

von Neumann, J. (1937). Some matrix inequalities and metrization of metric space. *Tomsk University Review*, **1**, 286-300. Reprinted in A.H. Taub (Ed.) (1962), *John von Neumann: Collected Works, Volume 4*. New York: Pergamon.

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

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Annals of Statistics*, **35**, 2654-2690.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.X. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 363-3410.

Yin, X., Li, B. and Cook, R.D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*. **99**, 1733-1757.

Table 1: Mean and standard error (in parenthesis) of the matrix distance in the simulations.

Design	n	p	BDR	BDR-I	BDR-II	SIR	DR	SR
1	100	6	0.05 (0.01)	0.08 (0.02)	0.07 (0.02)	1.60 (0.05)	0.39 (0.03)	0.99 (0.08)
2	100	6	0.59 (0.06)	0.54 (0.05)	1.60 (0.05)	1.39 (0.05)	1.29 (0.05)	1.33 (0.05)
3	100	6	0.26 (0.04)	0.23 (0.04)	0.40 (0.06)	2.59 (0.06)	0.56 (0.04)	1.33 (0.08)
4	100	6	0.51 (0.05)	0.38 (0.04)	0.51 (0.05)	1.59 (0.05)	1.51 (0.05)	1.53 (0.05)
5	400	10	0.01 (0.01)	0.12 (0.01)	0.01 (0.01)	0.33 (0.01)	0.46 (0.02)	0.21 (0.01)
1	500	20	0.01 (0.01)	0.05 (0.01)	0.01 (0.01)	1.84 (0.05)	0.24 (0.01)	0.09 (0.01)
2	500	20	0.22 (0.06)	0.44 (0.05)	1.12 (0.13)	1.55 (0.06)	1.46 (0.06)	1.71 (0.04)
3	500	20	0.13 (0.04)	0.14 (0.01)	0.08 (0.02)	3.61 (0.04)	0.27 (0.01)	0.21 (0.02)
4	500	20	0.38 (0.09)	0.36 (0.05)	0.26 (0.07)	1.93 (0.02)	1.69 (0.05)	1.87 (0.02)

Table 2: Proportion of the simulated data sets with inclusion probability (defined as the probability of being included in either latent sufficient predictor, P($\pi_{l1} = 1$ or $\pi_{l2} = 1$|y)) greater than 0.5 (x_{10} – x_{20} are omitted, a "$*$" indicates the truly important predictors).

Design	n	p	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	500	20	1.00*	1.00*	1.00*	0.00	1.00*	1.00*	0.00	0.02	0.02
2	500	20	1.00*	1.00*	1.00*	0.14	1.00*	1.00*	0.10	0.06	0.06
3	500	20	1.00*	1.00*	1.00*	0.04	0.98*	0.98*	0.02	0.04	0.10
4	500	20	0.94*	1.00*	0.96*	0.00	1.00*	1.00*	0.00	0.00	0.06
5	400	10	1.00*	1.00*	0.08	0.08	0.06	0.04	0.12	0.02	0.00

Table 3: Posterior summaries for the HIV data. The first two columns give the sample mean and standard deviation for each predictor at baseline, the remaining columns give the inclusion probability and 95% interval for the corresponding elements of A.

	Sample		ZDV monotherapy		Other treatments	
	mean	sd	In Prob	95% Int	In Prob	95% Int
Age (years)	35.3	8.7	0.05	(-0.06, 0.02)	0.12	(-0.16, 0.02)
Weight (kg)	75.1	13.3	0.07	(-0.15, 0.02)	0.07	(-0.02, 0.06)
Karnofsky score (0-100)	95.5	5.9	0.05	(-0.02, 0.07)	0.25	(-0.02, 0.25)
CD4 count (cells/mm^3)	350.5	118.6	1.00	(1.38, 2.35)	1.00	(1.59, 2.60)
CD8 count (cells/mm^3)	986.6	480.2	0.25	(-0.33, 0.02)	0.86	(-0.41, 0.00)
Hemophilia (Y/N)	0.08	0.28	0.25	(-0.33, 0.02)	0.27	(-0.25, 0.02)
Homosexual activity (Y/N)	0.66	0.47	0.05	(-0.03, 0.02)	0.08	(-0.03, 0.08)
Intravenous drug use (Y/N)	0.13	0.34	0.04	(-0.03, 0.02)	0.09	(-0.02, 0.13)
Non-white race (Y/N)	0.29	0.45	0.04	(-0.03, 0.02)	0.83	(-0.39, 0.01)
Male (Y/N)	0.83	0.38	0.05	(-0.02, 0.03)	0.12	(-0.16, 0.02)
Antiretroviral history (Y/N)	0.59	0.49	1.00	(-0.70, -0.26)	1.00	(-0.71, -0.29)
Symptomatic status (Y/N)	0.17	0.38	0.09	(-0.19, 0.02)	0.26	(-0.26, 0.02)

Figure 1: Plot of the conditional 0.025% and 0.975% quantiles (dashed lines) of $p(y|\lambda)$ for set of means μ_k (points) and cut points ϕ_k (vertical lines along the axis) with $\sigma_z = 0.2$ and $\sigma_y = 1$. The shade of the points is proportional to $p_k(\lambda)$ for $\lambda = 2$

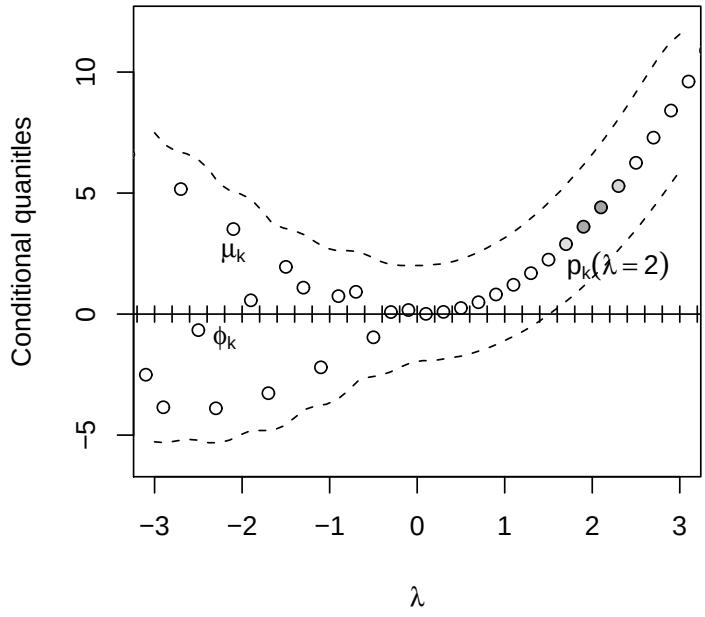
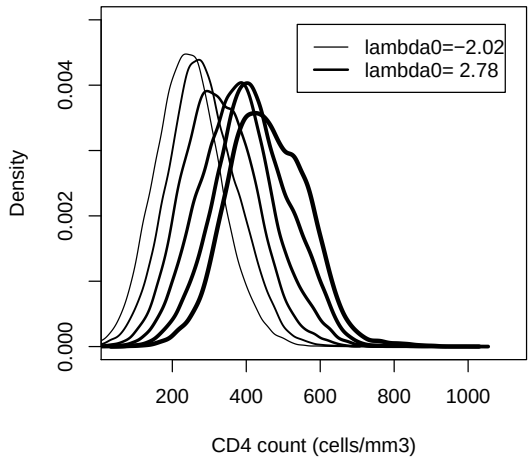
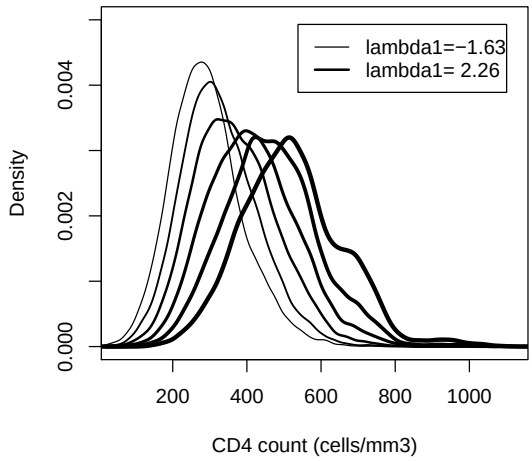


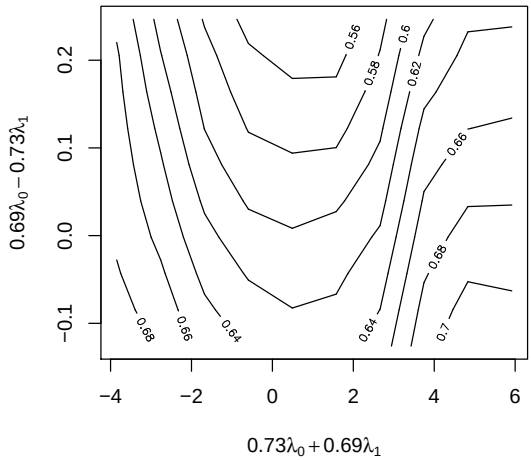
Figure 2: Summary of the CD4 analysis. Panels (a) and (b) plots of the estimated density of CD4 count 20 ± 5 weeks after baseline by the single index for ZDV monotherapy (Panel (a)) and the other treatments (Panel (b)). The values of the single index are equally space from -2.02 to 2.78 for ZDV monotherapy and -1.63 to 2.26 for other treatments. Panels (c) and (d) plot the probability that the response will be lower for ZDV monotherapy than the other treatments (left), and the difference (cells/mm^3) in the mean response for ZDV monotherapy and the other treatments (right), by the single index.



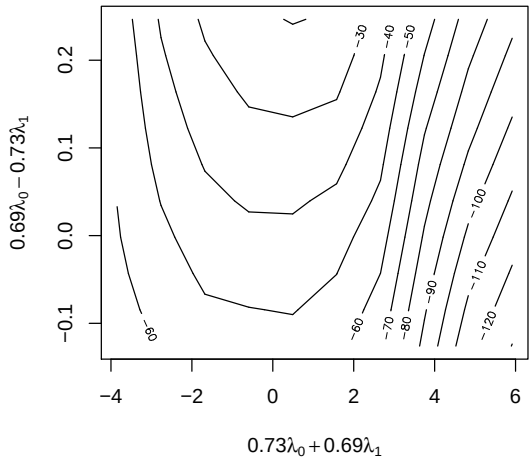
(a) Density for ZDV monotherapy



(b) Density for other treatments

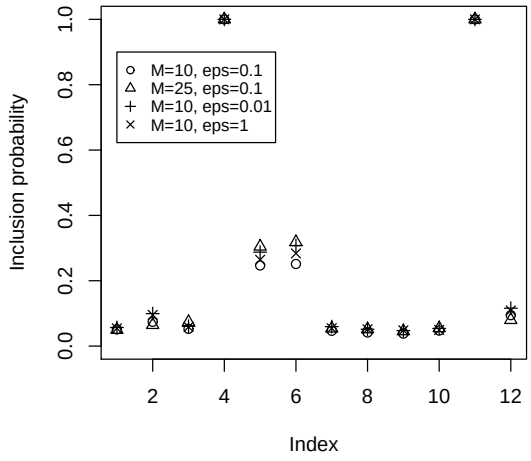


(c) Probability of lower CD4 count given ZDV monotherapy

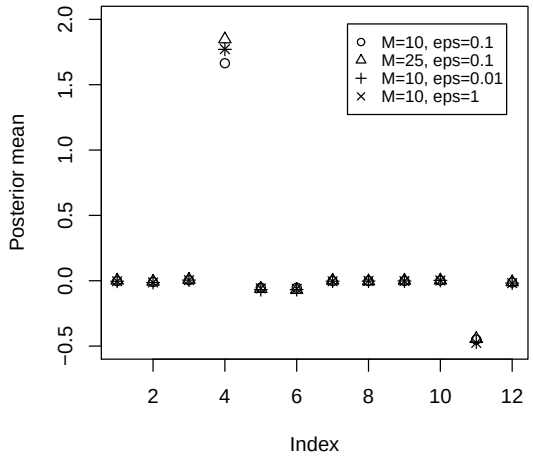


(d) Difference in mean responses

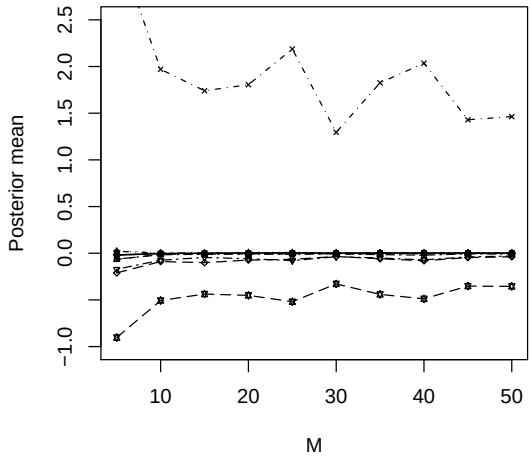
Figure 3: Sensitivity analysis. Inclusion probabilities and posterior means for A under different priors and number of terms M. Panels (c) and (d) assume $\epsilon = 0.1$



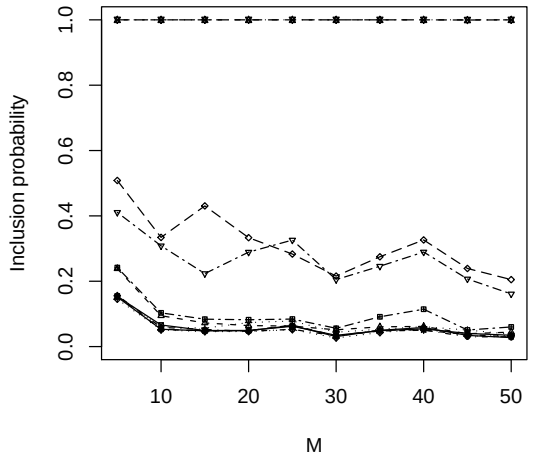
(a) Inclusion probabilities



(b) Posterior mean of A



(c) Inclusion probabilities by M



(d) Posterior means of A by M