Draft of the paper appeared in "Computational Statistics & Data Analysis, vol. 67, pp.105–114, 2013"

Entropy-Based Sliced Inverse Regression

Hideitsu Hino
University of Tsukuba.
1-1-1 Tennodai, Tsukuba, Ibaraki, 305–8573, Japan
Keigo Wakayama, and Noboru Murata
Waseda University, 3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan

Keywords: multiple kernel learning, entropy power inequality, discriminant analysis

Abstract

The importance of dimension reduction has been increasing according to the growth of the size of available data in many fields. An appropriate dimension reduction method of raw data helps to reduce computational time and to expose the intrinsic structure of complex data. Sliced inverse regression is a well-known dimension reduction method for regression, which assumes an elliptical distribution for the explanatory variable, and ingeniously reduces the problem of dimension reduction to a simple eigenvalue problem. Sliced inverse regression is based on the strong assumptions on the data distribution and the form of regression function, and there are a number of methods to relax or remove these assumptions to extend applicability of the inverse regression method. However, each method is known to have its drawbacks in either theoretically or empirically. To alleviate drawbacks in existing methods, a dimension reduction method for regression based on the notion of conditional entropy minimization is proposed. Using entropy as a measure of dispersion of data, a low dimensional subspace is estimated without assuming any specific distribution nor any regression function. The proposed method is shown to perform comparable or superior to conventional methods through experiments using artificial and real-world datasets.

1 Introduction

Dimension reduction is an important task in statistics, machine learning and data mining (Hastie et al., 2009; Bishop, 2006). During the past few decades, the importance of dimension reduction has grown as dimensions of available data have increased. When we deal with high dimensional data, an appropriate dimension reduction method of raw data helps to reduce computational time and storage resources. It also allows us to capture the intrinsic structure of target data (Roweis and Saul, 2000; Tenenbaum et al., 2000). Standard statistical methods often become unreliable for high dimensional data. This problem is usually referred to as the "curse of dimensionality" (Hastie et al., 2001), and it is of interest in many applications to reduce the dimension of the original data before constructing statistical models.

Dimension reduction methods can be divided into two categories: unsupervised and supervised methods. Typical unsupervised methods are principal component analysis (Jolliffe, 2002) and independent component analysis (Hyvärinen et al., 2001). On the other hand, Fisher's discriminant analysis (Fisher, 1936) and, for regression problems, *sliced inverse regression* (SIR, Li, 1991; Chen and Li, 1998) are used as supervised methods. In this paper, we consider a linear dimension reduction method for regression problems based on SIR, which aims at reducing the dimension of a vector-valued explanatory variable X while preserving its regression relation with a real-valued response variable Y. Li (1991) reduced the problem of dimension reduction to a simple eigenvalue problem, assuming that the distribution of the explanatory variable is elliptical (Owen and Rabinovitch, 1983) and the regression function is not symmetric. Under these assumptions, distributions of the data within any slices are summarized only by the within-slice means, and the dimension reduction subspace is identified by an eigenvalue problem. SIR is used in various literature, and for interesting applications of SIR, see, e.g., Wu and Lu (2004) and Wu and Lu (2007). However, it may happen that the data follow a non-elliptical distribution or the underlying regression function is symmetric, and in such cases, SIR fail to find dimension reduction subspaces (see section 4, model 1, for example). In order to avoid these problems, SAVE (Cook and S.Weisberg, 1991), DR (Li and Wang, 2007), and IRE (Cook and Ni, 2005) are developed by taking account of second order statistics. However, SAVE is not efficient in estimating monotone trends, and effectiveness of DR decays when the distribution of explanatory variable X deviates from a Gaussian distribution. IRE is a generalization of SIR and SAVE, however, in our experiments, it does not show satisfying performance (see section 4). To overcome limitations of these conventional methods, some dimension reduction methods for regression were recently proposed, e.g., Model-based SIR (MSIR) by Scrucca (2011), Least Square Dimension Reduction (LSDR) by Suzuki and Sugiyama (2010), Kernel Dimension Reduction (KDR) by Fukumizu et al. (2009), and dimension reduction for regression is still an area of active research. Among these recent studies, MSIR is a natural extension of SIR, which uses conditional distribution of the explanatory variable estimated using the sliced samples, while other two methods, LSDR and KDR, do not slice the samples. In this paper, we focus on slice-based methods, and propose a natural extension of SIR from a different perspective of MSIR. We propose a method based on the notion of *conditional entropy minimization* (Hino and Murata, 2010). Using entropy, which is estimated in non-parametric manner, as a measure of dispersion of data, a subspace on which the explanatory variables are projected is estimated without assuming the data distribution nor the form of regression function unlike other SIR-inspired methods. The proposed method is experimentally shown to perform comparable or superior to conventional methods.

The rest of this paper is organized as follows. Section 2 formulates the problem of dimension reduction for regression analysis. In section 3, a novel dimension reduction method based on the notion of conditional entropy minimization is proposed. Experimental results with artificial datasets and with various real-world datasets are given in section 4. The last section is devoted to concluding remarks.

2 Problem Formulation

Let X be a p-dimensional random explanatory variable and Y be a response variable. Realizations of X and Y are denoted by $\boldsymbol{x}$ and y, respectively. Linear dimension reduction seeks a set of linear combinations of X as $B^\top X$, where $B \in \mathbb{R}^{p\times q}$ $(q \leq p)$ is a *dimension reduction matrix* with column vectors $\boldsymbol{\beta}_i \in \mathbb{R}^p, i = 1, \dots, q$, such that Y depends on X only through linear combinations of $\{\boldsymbol{\beta}_1^\top X, \cdots, \boldsymbol{\beta}_q^\top X\}$. This can be formulated as $Y \perp X | B^\top X$, that is, Y and X are independent conditioned by $B^\top X$. When B satisfies this relation, its column space is called a *dimension reduction space*. Since we are interested in the column space of B, we assume B is an element of the Stiefel manifold $\mathbb{S}_q^p(\mathbb{R}) = \{B \in \mathbb{R}^{p\times q} | B^\top B = I_q\}$, where I_q is the $q \times q$ unit matrix. Under mild assumptions, the intersection of dimension reduction spaces is itself a dimension reduction space (Cook, 1998), and the intersection of all the dimension reduction spaces is called the central space. The main objective of dimension reduction is the statistical inference of the central space. More specifically, we assume a regression model

$$y = r(B^\top \boldsymbol{x}) + \varepsilon, \tag{1}$$

where r is an unknown regression function and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise term with variance σ^2. In many dimension reduction methods for regression including the one proposed in this paper, we neither specify nor estimate the regression function r, and our problem is to estimate B using a set of observed data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1,\ldots,n}$.

A notable feature of the inverse regression method is introduction of the distribution to the explanatory variable X. SIR estimation is based on the information provided by the inverse regression mean function $\mathrm{E}(X|Y)$. In practice, for a continuous response variable, the range of Y is sliced into L non-overlapping slices so that the numbers of observations in individual slices are approximately equal. See figure 1 for an illustrative example of the slicing. Assume the observed explanatory variables are standardized by $\tilde{\boldsymbol{x}}_i = \Sigma_{xx}^{-1/2}(\boldsymbol{x}_i - \boldsymbol{\mu})$, where Σ_{xx} and $\boldsymbol{\mu}$ are the sample covariance matrix and the sample mean of $\boldsymbol{x}$, respectively. We represent these slices by index sets $\mathcal{S}_l, l = 1, \ldots, L$ where $\cup_{l=1}^L \mathcal{S}_l = \{1, \ldots, n\}$, and $\mathcal{S}_l \cap \mathcal{S}_h = \emptyset,\ l \neq h$. We note that we use $\mathcal{S}_l$ to indicate both the index set and the corresponding subset of data. The cardinality of a set S is denoted by $|S|$. Then, variation on slice means, $\boldsymbol{\mu}_l = \frac{1}{|\mathcal{S}_l|}\sum_{i\in\mathcal{S}_l} \tilde{\boldsymbol{x}}_i,\ l = 1, \ldots, L$, yields the weighted covariance matrix $M = \frac{1}{L}\sum_{l=1}^L |\mathcal{S}_l| \boldsymbol{\mu}_l \boldsymbol{\mu}_l^\top$. Intuitively, since the observed data are transformed to have its center at the origin and have a unit covariance matrix, leading eigenvectors of M give intrinsic directions of the distribution of explanatory variable for accounting relationship between Y and X. Assuming the dimension q of the subspace is known and the distribution of $\tilde{\boldsymbol{x}}_i$ is spherical, it is shown in Li (1991) that the space spanned by leading q eigenvectors of M is a consistent estimate of the dimension reduction subspace. Thus, the dimension reduction matrix B is obtained from the eigen-decomposition of M followed by multiplication of $\Sigma_{xx}^{-1/2}$.

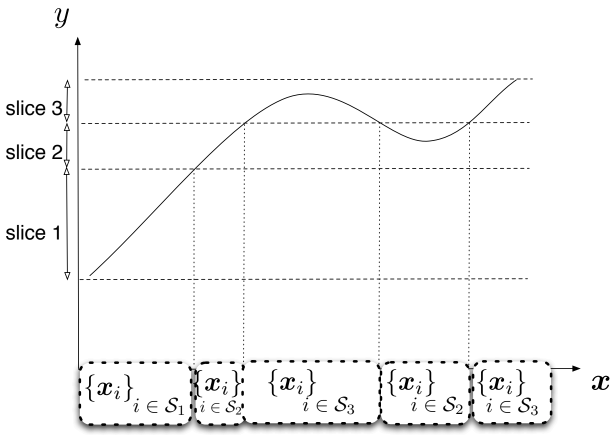

y
slice 3
slice 2
slice 1
$\{\boldsymbol{x}_i\}_{i \in \mathcal{S}_1}$
$\{\boldsymbol{x}_i\}_{i \in \mathcal{S}_2}$
$\{\boldsymbol{x}_i\}_{i \in \mathcal{S}_3}$
$\{\boldsymbol{x}_i\}_{i \in \mathcal{S}_2}$
$\{\boldsymbol{x}_i\}_{i \in \mathcal{S}_3}$
$\boldsymbol{x}$

Figure 1: Sliced region of response variable and corresponding explanatory variables.

3 Entropy-Based Sliced Inverse Regression

We propose an extension of SIR based on the notion of conditional entropy minimization (Hino and Murata, 2010), which is a supervised dimension reduction framework. When we refer to the term *entropy* in this paper, we mean the Shannon differential entropy (Cover and Thomas, 1991) for a random variable X defined as

$$H(X) = -\int f(\boldsymbol{x}) \log f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{2}$$

where f is the probability density function of X. The essential point of the extension is the fact that differential entropy can be seen as a generalization of variance. For example, the differential entropy of a Gaussian random variable with variance σ^2 is easily calculated as $H(X) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}$, where the variance and entropy are connected by the logarithmic function, which is a monotonically increasing function. By measuring the dispersion by differential entropy, which can be estimated in a non-parametric manner, we can take higher order statistics into account and we can estimate the dimension reduction subspace without assuming any specific distribution for X.

Example 1 (Figure 2) *This example illustrates the case that we can find a meaningful direction by conditional entropy minimization criterion.*

Suppose X in figure 1 is a two-dimensional explanatory variable. In figure 2, the explanatory variable follows the uniform distribution along with a directional vector $\boldsymbol{\beta}^* = (1/\sqrt{2}, 1/\sqrt{2})^\top$, and follows the Gaussian distribution with variance $\sigma^2 = 0.25$ along with the perpendicular direction $\hat{\boldsymbol{\beta}} = (1/\sqrt{2}, -1/\sqrt{2})^\top$. Consider the projections of the data in the second slice $\mathcal{S}_2$ by $\boldsymbol{\beta}^*$ and by $\hat{\boldsymbol{\beta}}$ onto one-dimensional axes. Then, data projected by $\boldsymbol{\beta}^*$ follow the uniform distribution

$$p_1(\boldsymbol{\beta}^{*\top}\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{\beta}^{*\top}\boldsymbol{x} \in [2/3, 1] \cup [8/3, 10/3] \\ 0, & otherwise \end{cases}, \tag{3}$$

where the variance of this distribution is $V(\boldsymbol{\beta}^{*\top}X) = 1.07$. On the other hand, data projected by $\hat{\boldsymbol{\beta}}$ follow the Gaussian distribution

$$p_2(\hat{\boldsymbol{\beta}}^\top\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{\boldsymbol{\beta}}^\top\boldsymbol{x} - \mu)^2}{2\sigma^2}\right), \tag{4}$$

where $V(\hat{\boldsymbol{\beta}}^\top X) = \sigma^2 = 0.25$ and μ is the mean on the axis.

The entropy of $\boldsymbol{\beta}^{*\top}X$ and $\hat{\boldsymbol{\beta}}^\top X$ are calculated as

$$\begin{aligned} H(\boldsymbol{\beta}^{*\top}X) &= E_{p_1}\left(-\log p_1(\boldsymbol{\beta}^{*\top}X)\right) = -\int_{t\in[2/3,1]\cup[8/3,10/3]} \log(1)\mathrm{d}t = 0, \\ H(\hat{\boldsymbol{\beta}}^\top X) &= E_{p_2}\left(-\log p_2(\hat{\boldsymbol{\beta}}^\top X)\right) = \frac{1}{2}\log 2\pi \cdot 0.25 + \frac{1}{2} = 0.73, \end{aligned}$$

where t is introduced for integrating $\boldsymbol{\beta}^{*\top}\boldsymbol{x}$. Since $H(\boldsymbol{\beta}^{*\top}X) < H(\hat{\boldsymbol{\beta}}^\top X)$ and $V(\boldsymbol{\beta}^{*\top}X) > V(\hat{\boldsymbol{\beta}}^\top X)$, when we find a dimension reduction subspace by minimizing the variance on the projected axis, we obtain $\hat{\boldsymbol{\beta}}$, while when we find it by minimizing the entropy, we obtain $\boldsymbol{\beta}^*$, which is the ground truth direction in this case.

The proposed algorithm minimizes the sum of conditional entropy of all sliced data $\{\boldsymbol{x}_i\}_{i\in\mathcal{S}_l}$, $l = 1, \ldots, L$. To minimize the entropy, we estimate the entropy of the projected data in a non-parametric manner. In this paper, we adopt the MeanNN entropy estimator $\hat{H}(\{\boldsymbol{x}_i\}_{i=1,\ldots,n})$

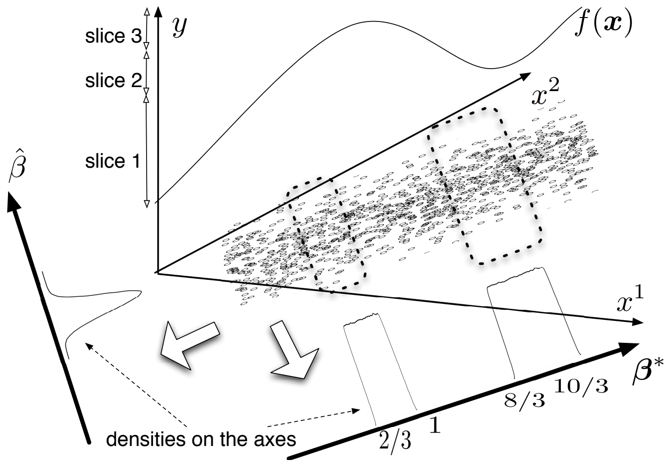


Figure 2: Projections on two axes by $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$ (Example 1).

proposed in Faivishevsky and Goldberger (2009) because of its stability, computational efficiency, and implementation simplicity:

$$\hat{H}(\{\boldsymbol{x}_i\}_{i=1,\ldots,n}) = \log(c_p) + \psi(n) + \frac{1}{n-1}\sum_{k=1}^{n-1}\left\{-\psi(k) + \frac{p}{n}\sum_{i\neq j}\log||\boldsymbol{x}_i - \boldsymbol{x}_j||\right\}, \tag{5}$$

where $c_p = \frac{\pi^{p/2}}{\Gamma(1+p/2)}$ is the volume of the q-dimensional unit ball, and ψ is the digamma function. Using this estimator, we estimate the entropy of projected data subsets in slices as $\hat{H}(\{B^\top \boldsymbol{x}_i\}_{i\in\mathcal{S}_l})$, which approximates *conditional entropies* of projected data $H(B^\top X|X \in \mathcal{S}_l)$. Then, setting the prior distribution for each slice $\mathcal{S}_l$ be $p(\mathcal{S}_l) = |\mathcal{S}_l|/n$, we define the minimization objective function $J(B)$ by the weighted sum of estimated entropies of projected data in each slice:

$$\begin{aligned} H(B^\top X|Y) &\simeq \sum_{l=1}^{L} p(\mathcal{S}_l) H(B^\top X|X \in \mathcal{S}_l) \qquad (6)\\ &\simeq \sum_{l=1}^{L} \frac{|\mathcal{S}_l|}{n}\hat{H}(\{B^\top \boldsymbol{x}_i\}_{i\in\mathcal{S}_l}) = J(B). \qquad (7)\end{aligned}$$

We minimize the objective function by the gradient descent method. The gradient matrix of the estimated conditional entropy is given by

$$\frac{\partial \hat{H}(\{B^\top \boldsymbol{x}_i\}_{i\in\mathcal{S}_l})}{\partial B^\top} = \frac{q}{|\mathcal{S}_l|(|\mathcal{S}_l|-1)}\sum_{\substack{i,j\in\mathcal{S}_l,\\ i\neq j}} \frac{B^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top}{\|B^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)\|^2}, \tag{8}$$

and the gradient matrix of the objective function is given by

$$\frac{\partial J(B)}{\partial B^\top} = \sum_{l=1}^{L}\frac{|\mathcal{S}_l|}{n}\frac{\partial \hat{H}(\{B^\top \boldsymbol{x}_i\}_{i\in\mathcal{S}_l})}{\partial B^\top}. \tag{9}$$

By updating B in the direction of $\frac{\partial J(B)}{\partial B^\top}$, B could deviate from the Stiefel manifold $\mathbb{S}_q^p(\mathbb{R})$. In the literature of independent component analysis, it is known that a matrix B is mapped to $\mathbb{S}_q^p(\mathbb{R})$

by a simple iterative algorithm, quasi-orthogonalization (Hyvärinen et al., 2001). We note that instead of projecting B after gradient steps, it is also possible to develop an algorithm that keeps B lay in the Stiefel manifold (Nishimori and Akaho, 2005; Fiori, 2005). We applied the latter approach in our preliminary study, but the simple gradient and projection method adopted in this paper showed better results in our case. The detail of the quasi-orthogonalization algorithm is shown in Appendix A. In Algorithm 1, we summarize the proposed *entropy-based sliced inverse regression* (ESIR) algorithm. In practice, we have to set the gradient parameter $\eta > 0$ appropriately. There are a lot of possibilities in finding an appropriate value of η. In this paper, we adopt a golden section search along the steepest descent direction. We set the stopping criterion be $|J(B_t) - J(B_{t-1})| < 10^{-8}$, where B_t and B_{t-1} are the estimated dimension reduction matrices at the t-th and $t-1$-th iterations, respectively.

Algorithm 1 ESIR: Entropy-based sliced inverse regression

```
**input**: a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1,\dots,n}$, gradient parameter $\eta > 0$, and reduced dimension $q$.
**initialize**: choose an initial matrix $B \in \mathbb{R}^{p \times q}$.
**while** : stopping criterion not met, **do**
    update $B$ by
        $B \leftarrow B - \eta \frac{\partial J(B)}{\partial B^\top}$.    (10)
    map $B$ to the Stiefel manifold $\mathbb{S}_q^p(\mathbb{R})$ by quasi-orthogonalization algorithm (Algorithm 2).
**end while**
```

4 Experimental Results

In this section, we show experimental results both on artificial and real-world data. We compare the proposed method to SIR, IRE, SAVE, DR, and MSIR, all of which have implementations openly available. In the experiments, we used R language (R Development Core Team, 2011) and its package `dr` for SIR, IRE, SAVE, DR, and package `msir` for MSIR. We also implemented the proposed ESIR algorithm using R.

4.1 Artificial Data

We show simulation results on the estimation of the ground truth projection matrix B^* of four models used in previous studies on SIR-based methods. In the same manner as Scrucca (2011), we evaluate the difference between the ground truth matrix B^* and estimated matrix $\hat{B}$ by the following measure:

$$\angle(\hat{B}, B^*) = \arcsin\left(\|\hat{B}(\hat{B}^\top\hat{B})^{-1}\hat{B}^\top - B^*(B^{*\top}B^*)^{-1}B^{*\top}\|_S\right),$$

which can be regarded as an angle between column spaces of B^* and $\hat{B}$. Here, the spectral norm $\|A\|_S$ is calculated by the maximum singular value of the matrix A.

In the following four models, an additive noise ε follows $\mathcal{N}(0, 0.1^2)$.

4.1.1 Model 1

We first consider the following single-index model with a symmetric regression function:

$$Y = (0.5 \cdot \boldsymbol{\beta}^\top X)^2 + \varepsilon, \tag{11}$$

where $\boldsymbol{\beta} = (1, -1, 0, \ldots, 0)^\top \in \mathbb{R}^{10}$. This model is a famous example in which the original SIR method cannot identify the ground truth matrix (vector). The explanatory variable X follows $\mathcal{N}(\mathbf{0}, I_{10})$, where $\mathbf{0} \in \mathbb{R}^{10}$.

4.1.2 Model 2

We consider the following two-dimensional model with polynomial regression function:

$$Y = \boldsymbol{\beta}_1^\top X + (\boldsymbol{\beta}_2^\top X)^2 + \varepsilon, \tag{12}$$

where $B = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2),\ \boldsymbol{\beta}_1 = (1, 0, \cdots, 0)^\top \in \mathbb{R}^{10}$ and $\boldsymbol{\beta}_2 = (0, 1, 0, \cdots, 0)^\top \in \mathbb{R}^{10}$. Theoretically, SIR can identify the first direction $\boldsymbol{\beta}_1$, but fail to identify $\boldsymbol{\beta}_2$. SAVE and DR take into account the second moment, and they can identify both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. The explanatory variable X follows $\mathcal{N}(\mathbf{0}, I_{10})$, where $\mathbf{0} \in \mathbb{R}^{10}$.

4.1.3 Model 3

We consider the following two-dimensional model with rational regression function:

$$Y = \frac{\boldsymbol{\beta}_1^\top X}{0.5 + (1.5 + \boldsymbol{\beta}_2^\top X)^2} + \varepsilon, \tag{13}$$

where $B = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2),\ \boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^\top, \boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)^\top$. This model is known as an example in which SAVE can not identify the ground truth dimension reduction matrix. The explanatory variable X follows $\mathcal{N}(\mathbf{0}, I_{10})$, where $\mathbf{0} \in \mathbb{R}^{10}$.

4.1.4 Model 4

Finally, we consider a model in which the explanatory variable has correlation. That is, X follows a 10-dimensional zero mean Gaussian distribution with correlation between X_i and X_j given by $\mathrm{cor}(X_i, X_j) = \rho^{|i-j|},\ i, j = 1, \ldots, 10$, and ρ is set to 0.05. The regression function is of the form

$$Y = 2\boldsymbol{\beta}^\top X + (\boldsymbol{\beta}^\top X)^2 + \varepsilon, \tag{14}$$

where $B = \boldsymbol{\beta} = (1, 1, 1, 0, \cdots, 0)^\top \in \mathbb{R}^{10}$.

4.1.5 Results and Discussion on the Artificial Data Experiments

Angles between the ground truth and estimated dimension reduction matrices are plotted in figures 3 (1-a), (2-a), (3-a) and (4-a), with increasing numbers of samples. The dimension reduction matrices are estimated using $n = 200, 300, \ldots, 1000$ samples. In every setting, 100 sets of i.i.d. samples are used to evaluate average performances of the dimension reduction methods.

From figures 3 (1-a),(2-a),(3-a), and (4-a), we see that ESIR performs better than the other methods in many cases. Figures 3 (1-b),(2-b),(3-b), and (4-b) show the averaged conditional estimated entropy in all slices, after projected onto estimated dimension reduction subspaces. From this result, we see that the performances of dimension reduction methods for regression are highly correlated with the magnitude of conditional entropy, which supports the validity of our proposed method.

For the sake of legibility, we did not show standard deviations of angles $\angle(\hat{B}, B^*)$ and entropies $J(\hat{B})$ in figure 3. Instead, in cases of $n = 200$ and $n = 1000$, we show averages of $\angle(\hat{B}, B^*)$ and entropies with standard deviations in parentheses in table 1 to table 4. The best methods that achieve the smallest $\angle(\hat{B}, B^*)$ are shown in boldface type when they are statistically significant via a t-test at significance level $\alpha = 0.05$.

Table 1: Model 1: Angles $\angle(\hat{B}, B^*)$ and estimated ones. Entropies of projected data by estimated are also listed.

	angle $\angle(\hat{B}, B^*)$		Conditional Entropy $J(\hat{B})$	
Sample Size	200	1000	200	1000
SIR	71.12 (16.64)	64.84(19.91)	1.29(0.07)	1.35(0.08)
IRE	76.25(11.75)	75.39 (15.84)	1.33(0.07)	1.38(0.05)
SAVE	18.44(5.21)	6.48(1.44)	0.99(0.10)	0.78(0.03)
DR	17.27 (4.82)	6.67(1.55)	0.97(0.10)	0.78 (0.03)
MSIR	10.77 (5.07)	2.88(0.89)	0.85(0.08)	0.71(0.02)
ESIR	**6.45** (2.29)	**2.00**(0.52)	0.69(0.07)	0.69(0.02)

Table 2: Model 2: Angles $\angle(\hat{B}, B^*)$ and estimated ones. Entropies of projected data by estimated are also listed.

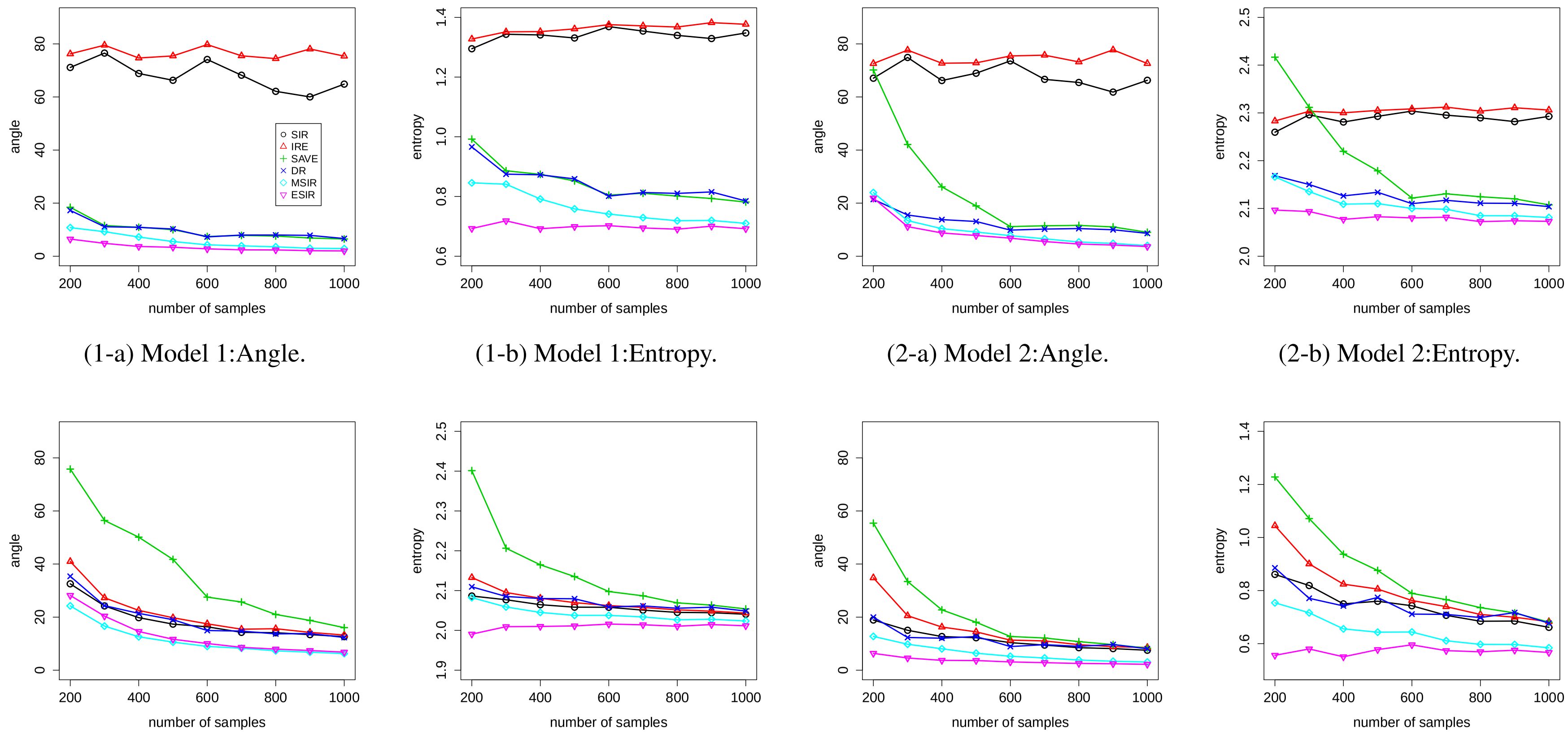
	angle $\angle(\hat{B}, B^*)$		Conditional Entropy $J(\hat{B})$	
Sample Size	200	1000	200	1000
SIR	67.07 (18.12)	66.29(18.88)	2.26(0.06)	2.29(0.05)
IRE	72.60(12.76)	72.61(16.74)	2.28(0.05)	2.31(0.04)
SAVE	70.17 (15.37)	9.01(2.15)	2.42(0.06)	2.11(0.03)
DR	21.42(5.91)	8.62(1.85)	2.17(0.06)	2.10(0.02)
MSIR	23.97 (21.55)	4.01(0.90)	2.17(0.08)	2.08(0.02)
ESIR	21.89(18.91)	**3.65**(0.72)	2.10(0.06)	2.07(0.02)

Table 3: Model 3: Angles $\angle(\hat{B}, B^*)$ and estimated ones. Entropies of projected data by estimated are also listed.

	angle $\angle(\hat{B}, B^*)$		Conditional Entropy $J(\hat{B})$	
Sample Size	200	1000	200	1000
SIR	32.54 (12.91)	12.65(3.52)	2.09(0.07)	2.04(0.03)
IRE	40.96(14.90)	13.26(3.98)	2.13(0.08)	2.04(0.03)
SAVE	75.77(11.08)	16.03(4.85)	2.40(0.09)	2.05(0.03)
DR	35.37 (14.07)	12.33(2.97)	2.11(0.07)	2.05(0.03)
MSIR	**24.23** (7.45)	**6.26**(1.67)	2.08(0.06)	2.02(0.02)
ESIR	28.17(11.25)	6.81(2.09)	1.99(0.06)	2.01(0.02)

Table 4: Model 4: Angles $\angle(\hat{B}, B^*)$ and estimated ones. Entropies of projected data by estimated are also listed.

	angle $\angle(\hat{B}, B^*)$		Conditional Entropy $J(\hat{B})$	
Sample Size	200	1000	200	1000
SIR	18.91 (6.02)	7.53(1.93)	0.86(0.15)	0.66(0.06)
IRE	34.81(16.80)	8.58(2.17)	1.04(0.19)	0.68(0.06)
SAVE	55.40(21.15)	8.30(2.72)	1.23(0.14)	0.68(0.06)
DR	19.98 (6.14)	8.08(2.08)	0.89(0.13)	0.67(0.05)
MSIR	12.74(4.62)	3.08(0.77)	0.75(0.11)	0.58(0.04)
ESIR	**6.35**(1.98)	**2.22**(0.57)	0.56(0.09)	0.57(0.04)



(1-a) Model 1:Angle.

(1-b) Model 1:Entropy.

(2-a) Model 2:Angle.

(2-b) Model 2:Entropy.

(3-a) Model 3:Angle.

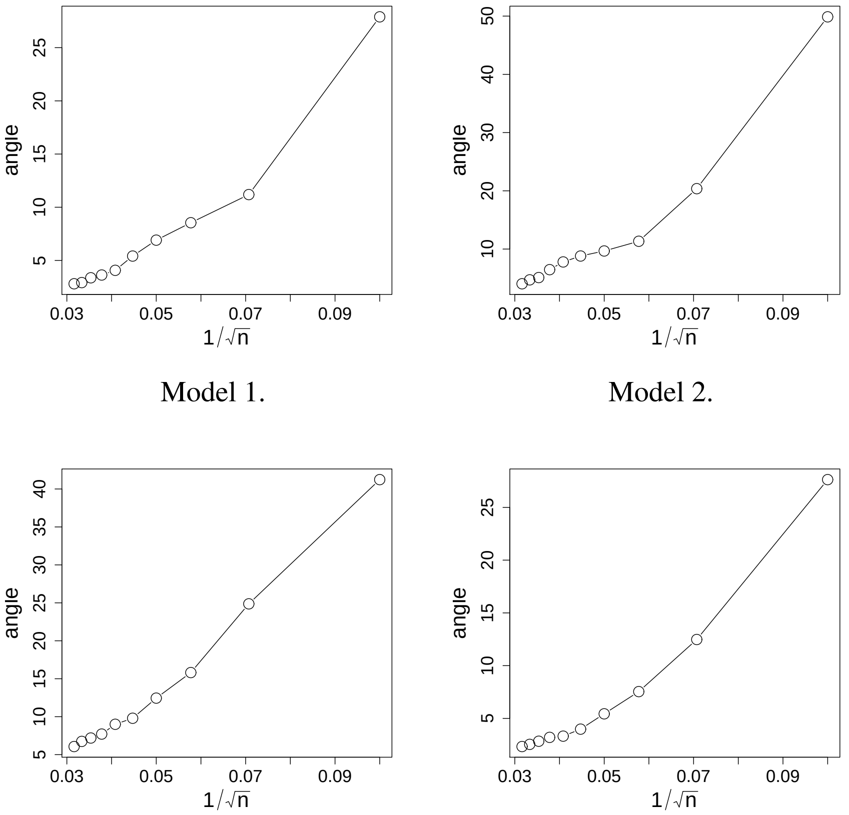
(3-b) Model 3:Entropy.

(4-a) Model 4:Angle.

(4-b) Model 4:Entropy.

Figure 3: Angles $\angle(\hat{B}, B^*)$, and conditional entropies $J(\hat{B})$ of the projected data by the estimated dimension reduction matrix.

Finally, asymptotic behaviours of statistical methods are important both theoretically and practically. Some SIR based methods are, based on their assumption on the distribution of the explanatory variable X, shown to have $\sqrt{n}$-consistency as estimators of the dimension reduction space. Since the propose method do not assume any distribution of X, it requires a different approach for investigating statistical properties, and we leave the theoretical investigation for our future work. In this paper, we show in figure 4 the averaged angles between the true and estimated subspaces by ESIR as functions of $1/\sqrt{n}$ for the four models. The averages are taken over 100 independent sample sets. Approximately linear relationships between angles and $1/\sqrt{n}$ are seen in these plots, and they suggest nearly $\sqrt{n}$-consistency of the proposed method.



Model 1.

Model 2.

Model 3.

Model 4.

Figure 4: Angles $\angle(\hat{B}, B^*)$ and $1/\sqrt{n}$.

4.2 Real-world Data

In this section, we apply dimension reduction methods based on SIR for 8 datasets. Datasets “airquality”, “auto-mpg”, “concrete”, and ”computer hardware” are obtained from the UCI machine learning repository (Murphy and Aha, 1994). Datasets “body fat” and “space_ga” are obtained from StatLib system (StatLib, 2012). Dataset “divorce rate” is obtained from the Vital Statistics of Japan yearly survey by the Ministry of Health, Labour and Welfare (Vital Statistics, 2012). The last dataset, “solar panel”, is a set of electricity data generated by solar panels in houses in Japan. All of the datasets are publicly available except the last one. In table 5, we show specifications of these datasets. In real-world data cases, there is no ground truth dimension reduction matrix. To evaluate the performance of dimension reduction methods, we fix a regression model to a simple linear regression of the form

Table 5: Dimensions of explanatory variables and numbers of samples of datasets.

data name	dimension	# of samples
airquality	5	111
auto-mpg	7	398
concrete	8	1030
computer hardware	9	209
body fat	14	252
space_ga	6	3107
divorce rate	12	97
solar panel	15	641

$$f(B^\top \boldsymbol{x}) = a_0 + a_1 \boldsymbol{\beta}_1^\top \boldsymbol{x} + \cdots + a_q \boldsymbol{\beta}_q^\top \boldsymbol{x}, \tag{15}$$

and trained the model using a training set $\{\hat{B}^\top \boldsymbol{x}_i, y_i\}_{i=1,\ldots,n_{tr}}$, where $\hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_q)$ is obtained by individual dimension reduction methods. Then, we compare the mean-squared error (MSE)

$$MSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n_{te}} (y_j - f(\hat{B}^\top \boldsymbol{x}_j))^2}, \tag{16}$$

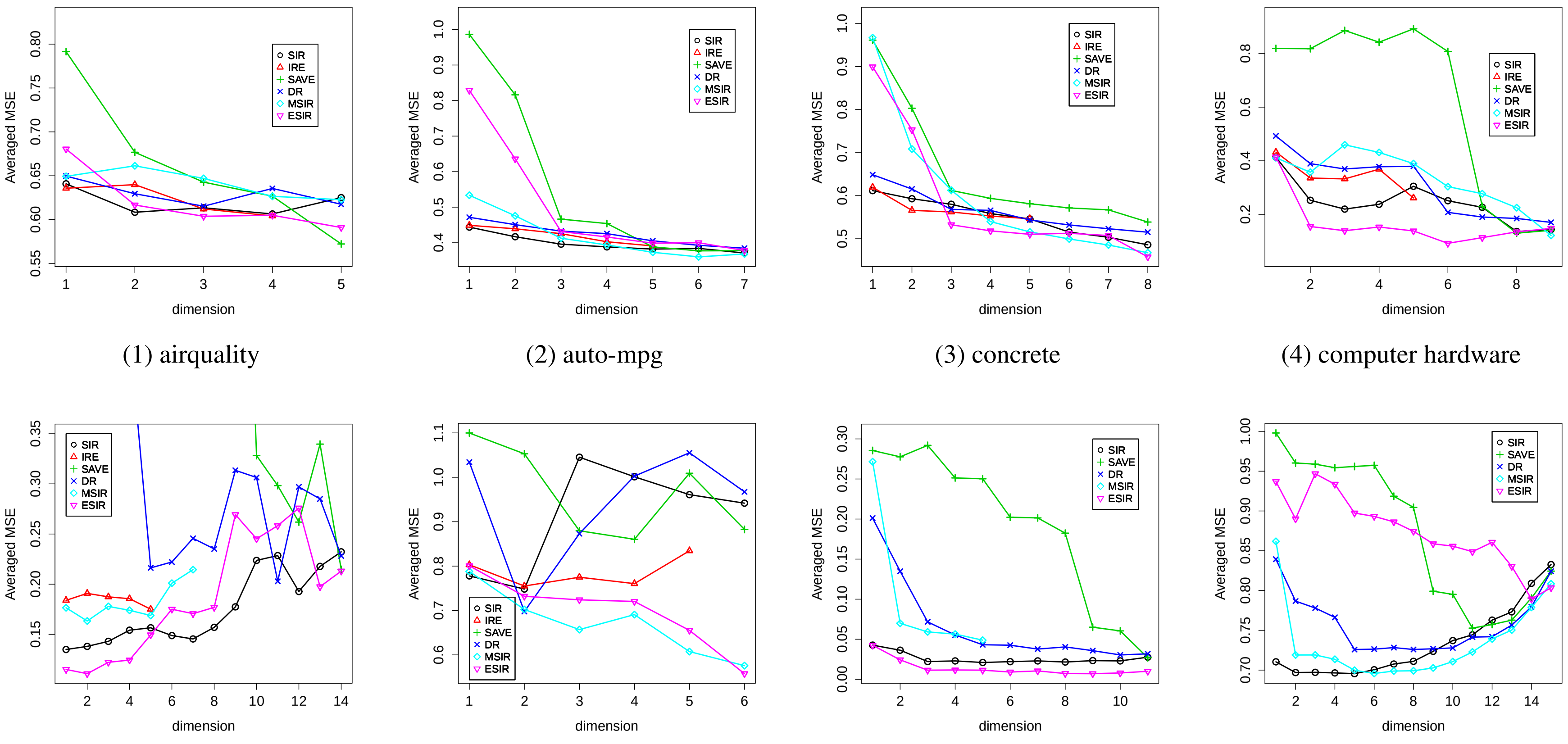
which is evaluated using a set of test data $\{\boldsymbol{x}_j, y_j\}_{j=1,\ldots,n_{te}}$. We report the averages of MSE estimated by 10-fold cross validation.

We note that for some datasets, IRE did not work because of numerical singularity. It is also noted that IRE and MSIR automatically limits the highest dimension of explanatory variables, hence the reduced dimension by IRE and MSIR for some datasets do not reach the maximum dimensions.

4.2.1 Results and Discussion on Real-world Datasets

There is no single best method that consistently outperforms the others, however, we can see that ESIR performs well for (3): "concrete", (4): "computer hardware", (5): "body fat", and (7): "divorce rate". Particularly, for "body fat" and "divorce rate" datasets, ESIR achieves the lowest MSE at only two- or three-dimensional subspaces. In the "body fat" dataset, we are supposed to predict the percentage of body fat using body density, age, weight, and so on. The estimation formula for the percentage of the body fat mainly depends on the body density, and other variables (e.g., age, weight, and chest circumference) seem to have minor contribution to the percentage of the body fat so much. In the "divorce rate" dataset, we are supposed to predict the divorce rate in a year using 12 observations such as the number of live births, number of deaths, number of infant deaths, number of divorces, number of marriages and so on, from 1898 to 1998 in Japan. It contains the number of divorces as one of the explanatory variables, hence it would be sufficient with only one or two dimensions for predicting the divorce rate. The proposed ESIR seems to be able to find these intrinsic low-dimensional subspaces from the given datasets.

On the other hand, ESIR does not perform well for “solar panel” dataset. Among 15 dimensions of explanatory variables of this dataset, 14 dimensions are categorical, and half of them (i.e., 7 dimensions) are binary variables. Because the Shannon differential entropy is defined for continuous probability distributions, it is understandable that the proposed method does not work well for this dataset.



(1) airquality (2) auto-mpg (3) concrete (4) computer hardware

(5) body fat (6) space_ga (7) divorce rate (8) solar panel

Figure 5: Averages of mean squared errors of linear regression using transformed explanatory variables by various methods.

5 Conclusion

In this paper, an extension of SIR is proposed based on the notion of conditional entropy minimization, which is a different perspective of model-based SIR (MSIR) by Scrucca (2011). The proposed dimension reduction method is based on non-parametric entropy estimation, hence it does not assume any specific distribution nor any regression function. It is shown to perform well for both artificial and real-world datasets.

A drawback of the proposed method would be possible sensitivity to the initial estimate of the dimension reduction matrix B, because the objective function of the method is non-convex in general. We adopted the result of SIR as an initial estimate of B for the proposed method. Indeed, we do not intend to replace the conventional methods. Instead, we propose to tune the estimated matrix B obtained through other methods by applying conditional entropy minimization. From our experimental results, we see that there is a strong correlation between the accuracy of estimation of the dimension reduction matrix and the value of the estimated conditional entropy.

In the experiment, we adopted the same heuristics as used in SIR for the choice of slices. It is also an open problem how to slice the range of response variable for the proposed method. In this paper, we only considered univariate response Y. SIR and related methods including our proposed method face with the curse of dimensionality when applied to regressions with multivariate responses. There are some works on how to slice multivariate responses to extend SIR based method for multivariate response regressions (Aragon, 1997; Setodji and Cook, 2004). Combining such methods with ESIR would further broaden the applicability of SIR based methods. In practice, we usually do not have any prior knowledge about q, the dimension of the dimension reduction subspace. Some prior works on inverse regression entail statistical tests for determining appropriate dimensions. At this moment, we suggest using the same q determined by such methods in practical circumstances. A method for determining an appropriate dimension based on the notion of entropy should be explored. Statistical properties of the proposed method seeing as an estimator for dimension reduction subspaces are yet to be studied. Most of statistical properties such as consistency, unbiasedness, and identifiability must be closely connected to both properties of the entropy estimator and the way to slice the given dataset, and one of our important future works is investigation of these properties.

Acknowledgement

The authors would like to express our special thanks to the editor, the associate editor and reviewers whose comments led to valuable improvements of the manuscripts. The authors are grateful to S. Weisberg and L. Scrucca for providing implementation of SIR, IRE, SAVE, DR, and MSIR. Part of this work was supported by "Development of Large Scale Energy Storage System with High-safety and Cost-competitiveness" from NEDO, Japan.

Appendix A: Quasi-orthogonalization

The quasi-orthogonalization of a matrix B is realized by iterating the three steps shown in algorithm 2 until convergence.

This procedure for quasi-orthogonalization is validated as follows (Hyvärinen et al., 2001). Let $B^\top B = EDE^\top$ be the eigenvalue decomposition of the symmetric matrix $B^\top B$, where $E \in \mathbb{R}^{p \times p}$ is an orthogonal matrix and D is a diagonal matrix with eigenvalues $\{d_i\}_{i=1}^{p}$ of

Algorithm 2 quasi-orthogonalization of matrix B

```
while :stopping criterion not met, do
    step 1: divide $B$ by square root of the largest eigenvalue of $B^\top B$.
    step 2: $B \leftarrow \frac{3}{2}B - \frac{1}{2}BB^\top B$.
    step 3: normalize the norm of each column of $B$ to 1.
end while
```

$B^\top B$. Then, by step 2 of the above procedure, $B^\top B$ is modified as

$$\begin{aligned} B^\top B &\mapsto \frac{1}{4}(3B - BB^\top B)^\top (3B - BB^\top B) \\ &= \frac{1}{4}E\left(9D - 6D^2 + D^3\right)E^\top. \end{aligned}$$

Noting that $d_i \in [0,1]$ because the maximum eigenvalue of the matrix $B^\top B$ is normalized to one in step 1, the eigenvalues of $B^\top B$ after this transformation become

$$h(d_i) = \frac{1}{4}(9d_i - 6d_i^2 + d_i^3), \quad i = 1, \ldots, p.$$

Because $h(d_i) - d_i = \frac{d_i}{4}\{(d_i - 3)^2 - 4\} \geq 0$, eigenvalues of $B^\top B$ converge to 1 by iterating those three steps. In actual experiments, we iterate these three steps $2 \times p$ times to obtain an approximately orthogonalized matrix.

References

Aragon, Y., 1997. A Gaussian implementation of multivariate sliced inverse regression. Computational Statistics 12, 355–372.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Chen, C.-H., Li, K.-C., 1998. Can SIR be as popular as multiple linear regression? Statistica Sinica 8, 289–316.

Cook, R. D., 1998. Regression Graphics. Wiley-Interscience.

Cook, R. D., S.Weisberg, 1991. Sliced inverse regression for dimension reduction: Comment. Journal of the American Statistical Association 86 (414), 328–332.

Cook, R. D., Ni, L., 2005. Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach. Journal of the American Statistical Association 100 (470), 410–428.

Cover, T. M., Thomas, J. A., 1991. Elements of information theory. John Wiley and Sons, Inc.

Faivishevsky, L., Goldberger, J., 2009. ICA based on a smooth estimation of the differential entropy. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21, 433–440.

Fiori, S., 2005. Formulation and integration of learning differential equations on the Stiefel manifold. IEEE Transactions on Neural Networks. 16(6), 1697–1701.

Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. Annals Eugen. 7, 179–188.

Fukumizu, K., Bach, F. R., Jordan, M. I., 2009. Kernel dimension reduction in regression. Ann. Stat. 37, 1871–1905.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning — Data Mining, Inference, and Prediction, 1st Edition. Springer.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning — Data Mining, Inference, and Prediction, 2nd Edition. Springer.

Hino, H., Murata, N., 2010. A conditional entropy minimization criterion for dimensionality reduction and multiple kernel learning. Neural Computation 22(11), 2887–2923.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. J. Wiley, New York.

Jolliffe, I., 2002. Principal component analysis. Springer Verlag, New York.

Kent, J.-T., 1991. Comment on "sliced inverse regression for dimension reduction". by K.C.Li, Journal of the American Statistical Association 86 (414), 316–342.

Li, B., Wang, S., 2007. On directional regression for dimension reduction. Journal of the American Statistical Association 102, 997–1008.

Li, K.-C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86 (414), 316–327.

Murphy, P., Aha, D., 1994. UCI Repository of machine learning databases. Tech. rep., University of California, Department of Information and Computer Science, Irvine, CA, US.

Nishimori, Y., Akaho, S., 2005. Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. Neurocomputing 67, 106–135.

Owen, J., Rabinovitch, R., 1983. On the class of elliptical distributions and their applications to the theory of portfolio choice. Journal of Finance 38 (3), 745–52.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org/`

Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding 290 (5500), 2323–2326.

Scrucca, L., 2011. Model-based SIR for dimension reduction. Computational Statistics & Data Analysis 55 (11), 3010–3026.

Setodji, C. M., Cook, R. D., 2004. K-Means inverse regression. Technometrics 46(4), 421–429.

StatLib: Data, Software and News from the Statistics Community, the Department of Statistics at Carnegie Mellon University.
URL `http://lib.stat.cmu.edu/index.php/`

Suzuki, T., Sugiyama, M., 2010. Sufficient dimension reduction via squared-loss mutual information estimation. In: Teh, Y. W., Tiggerington, M. (Eds.), Proc. of the Thirteenth International Conference on Artificial Intelligence and Statistics 9, 804–811.

Tenenbaum, J., de Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Wu, H. M., Lu, H. H.-S., 2004. Supervised Motion Segmentation by Spatial-Frequential Analysis and Dynamic Sliced Inverse Regression. Statistica Sinica 14, 413–430.

Wu, H. M., Lu, H. H.-S., 2007. Iterative Sliced Inverse Regression for Segmentation of Ultrasound and MR Images. Pattern Recognition 40(12), 3492–3502.

The Yearly Vital Statistics, the Ministry of Health, Labour and Welfare, Japan.
URL http://www1.mhlw.go.jp/toukei/kjd100_8/index.html