Localized Sliced Inverse Regression

Qiang Wu, Feng Liang, and Sayan Mukherjee[*]

We develop an extension of sliced inverse regression (SIR) that we call localized sliced inverse regression (LSIR). This method allows for supervised dimension reduction on nonlinear subspaces and alleviates the issue of degenerate solutions in the classical SIR method. We introduce a simple algorithm that implements this method. The method is also extended to the semisupervised setting where one is given labeled and unlabeled data. We illustrate the utility of the method on real and simulated data. We also note that our approach can interpolated between SIR and principle components analysis (PCA) depending on parameter settings.

Key Words: dimension reduction, sliced inverse regression, localization, semi-supervised learning

[*] Qiang Wu is a Postdoctoral Research Associate in the Departments of Statistical Science and Computer Science and the Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708-0251, U.S.A. (Email: qiang@stat.duke.edu). Feng Liang is an Assistant Professor in the Department of Statistics, University of Illinois at Urbana-Champaign, IL 61820, U.S.A. (Email: liangf@uiuc.edu). Sayan Mukherjee is an Assistant Professor in the Departments of Statistical Science and Computer Science and the Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708-0251, U.S.A. (Email:sayan@stat.duke.edu).

1 Introduction

The importance of dimension reduction for predictive modeling and visualization has a long and central role in statistical graphics and computation (Adcock, 1878; Edgeworth, 1884; Fisher, 1922; Young, 1941). In the modern context of high-dimensional data analysis this perspective posits that the functional dependence between a response variable y and a large set of explanatory variables $x \in \mathbb{R}^p$ is driven by a low dimensional subspace of the p variables. Characterizing this predictive subspace, supervised dimension reduction, requires both the response and explanatory variables. This problem in the context of linear subspaces or Euclidean geometry has been explored by a variety of statistical models such as sliced inverse regression (SIR (Li, 1991)), sliced average variance estimation (SAVE, (Cook and Weisberg, 1991)), principal Hessian directions (pHd, (Li, 1992)), (conditional) minimum average variance estimation (MAVE (Xia et al., 2002)), and extensions to these approaches.

In machine learning community research on nonlinear dimension reduction in the spirit of Young (1941) has been developed of late. This has led to a variety of manifold learning algorithms such as isometric feature mapping (ISOMAP, (Tenenbaum et al., 2000)), local linear embedding (LLE, (Roweis and Saul, 2000)), Hessian Eigenmaps (Donoho and Grimes, 2003), and Laplacian Eigenmaps (Belkin and Niyogi, 2003). Two key differences exist between the paradigm explored in this approach and that of supervised dimension reduction. The first difference is that the above methods are unsupervised in that the algorithms take into account only the explanatory variables. This issue can be addressed by extending the unsupervised algorithms to use the label or response data (Globerson and Roweis, 2006). The bigger problem is that these manifold learning algorithms do not operate on the space of the explanatory variables and hence do not provide a predictive submanifold onto which the data should be projected. These methods are based on embedding the observed data onto a graph and then using spectral properties of the embedded graph for dimension reduction. The key observation in all of these manifold algorithms is that metrics must be local and properties that hold in an ambient Euclidean space are true locally on smooth manifolds.

This suggests that the use of local information in supervised dimension reduction methods may be of use to extend methods for dimension reduction to the setting of nonlinear subspaces and submanifolds of the ambient space. In the context of mixture modeling for classification two approaches have been developed (Hastie and Tibshirani, 1996; Sugiyam, 2007).

In this paper we extend SIR by taking into account the local structure of the explanatory variables conditioned on the response variable. This localized variant of SIR, LSIR, can be used for classification as well as regression applications. Though the predictive directions obtained by LSIR are linear ones, they coded nonlinear information. Another advantage of our approach is that ancillary unlabeled data can be easily added to the dimension reduction analysis – semi-supervised learning.

The paper is arranged as follows. Localized sliced inverse regression is introduced in Section 2 for continuous and categorical response variables. Extensions are discussed in Section 3. The utility of the method with respect to predictive accuracy as well as exploratory data analysis via visualization is demonstrated on a variety of simulated and real data in Sections 4 and 5. We close with discussions in Section 6.

2 Local SIR

In this section we start with a brief review of sliced inverse regression. We remark that the failure of SIR in some situations is due its inability to consider local structure. This motivates a generalization of SIR, local SIR, that incorporates localization ideas from manifold learning into the SIR framework. We close by relating our extension to some classical methods for dimension reduction.

2.1 Sliced inverse regression

Assume the functional dependence between a response variable Y and an explanatory variable $X \in \mathbb{R}^p$ is given by

$$Y = f(\beta_1' X, \ldots, \beta_L' X, \varepsilon), \tag{1}$$

where $\{\beta_1, ..., \beta_L\}$ are unknown orthogonal vectors in $\mathbb{R}^p$ and ε is noise independent of X. Let B denote the L-dimensional subspace spanned by $\{\beta_1, ..., \beta_L\}$. $P_B X$, where P_B denotes the projection operator onto space B, provides a sufficient summary of the information in X relevant to Y, $Y \perp\!\!\!\perp X | P_B X$. Estimating B becomes the central problem in supervised dimension reduction. Though we define B here via a model assumption (1), a general definition based on conditional independence between Y and X given $P_B X$ can be found in Cook and Yin (2001). Following Cook and Yin (2001), we refer to B as the dimension reduction (d.r.) subspace and $\{\beta_1, ..., \beta_L\}$ as the d.r. directions.

The Slice inverse regression (SIR) model was introduced in Duan and Li (1991) and Li (1991) to estimate the d.r. directions. Consider a simple case where X has an identity covariance matrix. The conditional expectation $\mathbb{E}(X|Y=y)$ is a curve in $\mathbb{R}^p$ on which y varies. It is called the inverse regression curve since the position of X and Y are switched as compared to the classical regression setting, $\mathbb{E}(Y|X=x)$. It is shown in Li (1991) that the inverse regression curve is contained in the d.r. space B under some mild assumptions. According to this result the d.r. directions $\{\beta_1, ..., \beta_L\}$ correspond to eigenvectors with nonzero eigenvalues of the covariance matrix $\Gamma = \mathrm{Cov}(\mathbb{E}(X|Y))$. In general when the covariance matrix of X is Σ, the d.r. directions can be obtained by solving a generalized eigen decomposition problem

$$\Gamma\beta = \lambda\Sigma\beta.$$

The following simple algorithm implements SIR. Given a set of observations $\{(x_i, y_i)\}_{i=1}^n$:

1. Compute an empirical estimate of Σ,

   $$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)(x_i - m)^T$$

   where $m = \frac{1}{n}\sum_{i=1}^n x_i$ is the sample mean.

2. Divide the samples into H groups (or *slices*) $G_1, \ldots, G_H$ according to the value of y. Compute an empirical estimate of Γ,

   $$\hat{\Gamma} = \sum_{i=1}^{H}\frac{n_h}{n}(m_h - m)(m_h - m)^T.$$

   where

   $$m_h = \frac{1}{n_h}\sum_{j \in G_h} x_j$$

   is the sample mean for group G_h with n_h being the group size.

3. Estimate the d.r. directions β by solving a generalized eigen decomposition problem

$$\hat{\Gamma}\beta = \lambda\hat{\Sigma}\beta. \tag{2}$$

When y takes categorical values as in classification problems, it is natural to divide the data into different groups by their group labels. In the case of two groups SIR is equivalent to Fisher discriminant analysis (FDA, which is also known as linear discriminant analysis in some literatures).

SIR has been widely used for dimension reduction with success in practice. However, it has some known shortcomings. For example, it is easy to construct a function f such that $\mathbb{E}(X|Y = y) = 0$ and in this case SIR will fail to retrieve any useful directions (Cook and Weisberg, 1991). The degeneracy of SIR also restricts its use in binary classification problems since only one direction can be obtained. The failure of SIR in these scenario is partly due to the fact that the algorithm uses only the mean in each slice, $\mathbb{E}(X|Y = y)$, to summarize information in the slice. For nonlinear structures this is clearly not enough information. Generalizations of SIR such as SAVE (Cook and Weisberg, 1991), SIR-II (Li, 2000), and covariance inverse regression estimation (CIRE, Cook and Ni (2006)) address this issue by the second moment information on the conditional distribution of X given Y. It may not be enough however to use moments or global summary statistics to characterize the information in each slice. For example, analogous to the *multimodal* situation considered by Sugiyam (2007), the data in a slice may form two clusters for which global statistics such as moments would not provide a good description of the data. The cluster centers would be good summary statistics in this case. We now propose a generalization of SIR that uses local statistics based on the local structure of the explanatory variables in each slice.

2.2 Localization

A key principle in manifold learning is that Euclidean structure around a data point in $\mathbb{R}^p$ is only meaningful locally. Under this principle computing a global average m_h for a slice is not meaningful since some of the observations in a slice may be far apart in the ambient space. Instead we should consider local averages. Localized SIR (LSIR) implements this idea.

We first provide an intuition of the method. Without loss of generality we consider a slice and the transformation of the data in the slice by its empirical covariance. In the original SIR method we would shift the all the transformed data points by the corresponding group average and then perform a spectral decomposition on the covariance matrix of the transformed and shifted data to identify the SIR directions. The rational behind this approach is that if a direction does not differentiate different groups well, the group means projected onto that direction would be very close and therefore the variance of the transformed data will be small in that direction. A natural way to incorporate localization idea into this approach is to shift each transformed data point to the average of a local neighborhood instead of the average of its global neighborhood (i.e., the whole group). In manifold learning, local neighborhoods are often defined by the k nearest neighbors, k-NN, of a point. Note that the neighborhood selection in LSIR takes into account locality of points in the ambient space as well as information about the response variable due to slicing.

Recall that in SIR the group average m_h is used in estimating $\Gamma = \mathrm{Cov}(\mathbb{E}(X|Y))$

and the estimate $\hat{\Gamma}$ is equivalent to the sample covariance of a data set $\{m_i\}_{i=1}^n$ with $m_i = m_h$, the average of the group G_h to which x_i belongs. In our LSIR algorithm, we set m_i equal to a local average of observations in group G_h near x_i. We then use the corresponding sample covariance matrix to replace $\hat{\Gamma}$ in equation (2).

The following algorithm implements LSIR:

1. Compute $\hat{\Sigma}$ as in SIR.

2. Divide the samples into H groups as in SIR. For each sample (x_i, y_i) compute

$$m_{i,\text{loc}} = \frac{1}{k} \sum_{j \in s_j} x_j,$$

   where h indexes the group and for observations $i \in G_h$,

$$s_i = \{j : x_j \text{ belongs to the } k \text{ nearest neighbors of } x_i \text{ in } G_h\}.$$

   Then we compute a localized version of Γ by

$$\hat{\Gamma}_{\text{loc}} = \frac{1}{n} \sum_{i=1}^n (m_{i,\text{loc}} - m)(m_{i,\text{loc}} - m)^T.$$

3. Solve the generalized eigen decomposition problem

$$\hat{\Gamma}_{\text{loc}} \beta = \lambda \hat{\Sigma} \beta. \tag{3}$$

The neighborhood size k in LSIR is a tuning parameter specified by users. When k is large enough, say, larger than the size of any group, then $\hat{\Gamma}_{\text{loc}}$ is the same as $\hat{\Gamma}$ and LSIR recovers all SIR directions. On the other hand, when k is small, say, equal to 1, then $\hat{\Gamma}_{\text{loc}} = \hat{\Sigma}$ and LSIR keeps all p dimensions. A regularized version of LSIR, which we introduce in the next section, with $k = 1$ is empirically equivalent to principal components analysis, see Appendix A. In this light by varying k LSIR bridges between PCA and SIR and can be expected to retrieve directions lost by SIR due to degeneracy.

For classification problems LSIR becomes a localized version of FDA. Suppose the number of classes is C, then the estimate $\hat{\Gamma}$ from the original FDA is of rank at most $C - 1$, which means FDA can only estimate at most $C - 1$ directions. This is why FDA is seldom used for binary classification problems where $C = 2$. In LSIR we use more than the centers of the two classes to describe the data. Mathematically this is reflected by the increase of the rank of $\hat{\Gamma}_{\text{loc}}$ which is no longer bounded by C and hence produces more directions. Moreover, if for some classes the data is composed of several sub-clusters, LSIR can automatically identify these sub-cluster structures. We will show in one of our examples, this property of LSIR is very useful in data analysis such as cancer subtype discovery using genomic data.

2.3 Connection to Existing Work

The idea of localization has been introduced in dimension reduction for classification problems. For example, local discriminant information (LDI) introduced by Hastie and Tibshirani (1996). In LDI, local information is used to compute the between-group covariance matrix Γ_i over a nearest neighborhood at every data point x_i and then estimate the d.r. directions by the top eigenvector of the averaged between-group matrix $\frac{1}{n}\sum_{i=1}^{n}\Gamma_i$. Local Fisher discriminant analysis (LFDA) introduced by Sugiyam (2007) can be regarded as an improvement of LDI with the within-class covariance matrix also localized.

Compared to these two approaches, LSIR utilizes the local information directly at the point level which consequently affects the covariance matrix. One advantage of this simple localization is computation. For example, for a problem of C classes, LDI needs to compute $(n \times C)$ local mean points at each location x_i and n between-group covariance matrices, while LSIR computes only n local mean points and one covariance matrix.

Due to this simple localization at the point level, LSIR can be easily extended to handle unlabeled data in semi-supervised learning as explained in the next section. Such an extension is less straightforward for the other two approaches that operate on the covariance matrices instead of data points.

3 Extensions of LSIR

In this section we discuss some extensions of our LSIR algorithm.

3.1 Regularization

When the matrix $\hat{\Sigma}$ is singular or has a very large condition number, which is common in high-dimensional problems, generalized eigen decomposition problems (3) are unstable. Regularization techniques are often introduced to address this issue. For LSIR we adopt the following regularization:

$$\hat{\Gamma}_{\text{loc}}\beta = \lambda(\hat{\Sigma} + s)\beta \tag{4}$$

where $s > 0$ is a regularization parameter. A similar adjustment has also been proposed in Hastie and Tibshirani (1996). In practice, we recommend trying different values of s coupled with a sensitivity analysis of s. A more data driven way to select s is to introduce a criteria measuring the goodness of dimension reduction, such as the ratio of between group variance and within group variance, then use cross-validation to choose s; see e.g. Zhong et al. (2005).

3.2 Localization methods

In Section 2.2 we have suggested using k-nearest neighbors to localize data points. Other choices for localization include weighted averages of data using kernels of various bandwidths. This method, given a positive function decreasing on $\mathbb{R}^+$, computes for each sample (x_i, y_i) the local mean as

$$\hat{m}_{i,\text{loc}} = \frac{\sum_{j \in G_h} x_j W(\|x_j - x_i\|)}{\sum_{j \in G_h} W(\|x_j - x_i\|)}$$

where G_h is the group containing the sample (x_i, y_i). Examples of the function W include the Gaussian kernel and the flat kernel

$$\begin{aligned} W(t) &= e^{-t^2/\sigma^2} \\ W(t) &= 1, \text{ if } t \le r, \quad 0, \text{ if } t > r. \end{aligned}$$

A common localization approach in manifold learning is to truncate a smooth kernel by multiplying it by a flat kernel.

For any kernel, there is a bandwidth parameter (σ or r), which plays the same role as the parameter k in k-NN. Sensitivity analysis or cross-validation is recommended for the selection of these parameters.

3.3 Semi-supervised learning

In semi-supervised learning some data are labeled – observations of both the response as well as the explanatory variables are available – and some of the data are unlabeled – only observations of the explanatory are available. There are two obvious suboptimal approaches for dimension reduction in this setting: a) ignore the response variable and apply PCA to the entire data; b) ignore the unlabeled data and apply supervised dimension reduction methods such as SIR to this subset. Either method ignores data.

The point of semi-supervised learning is the incorporation of information from the unlabeled data into supervised analysis of the labeled data. LSIR can be easily modified to take the unlabeled data into consideration. Since we do not know the response variables of the unlabeled data we assume that they can take any value and so we add the unlabeled data into all slices. This defines a neighborhood s_i as the following: any point in the k-NN of x_i belongs to s_i if it is unlabeled or if it is labeled and belongs to the same slice as x_i. To reduce the influence of the unlabeled data one can put different weights between labeled and unlabeled points in calculating $m_{i,\text{loc}}$.

In the next section we demonstrate that our algorithm performs very well when just a small fraction of points in a data set are labeled and other dimension reduction methods fail to retrieve the relevant directions.

4 Simulations

In this section we apply LSIR to several synthetic data sets. We normalize each of the p dimensions to have unit variance, so the noise-to-signal ratio is about $(1 - L/p)$ where

L denotes the number of relevant dimensions. We compare the performance of LSIR to other dimension reduction methods such as SIR, SAVE, pHd, and LFDA.

We introduce the following metric to measure the accuracy in estimating the d.r. space B. Let $\hat{B} = (\hat{\beta}_1, \cdots, \hat{\beta}_L)$ denote an estimate of B, the columns $\hat{\beta}_i$ of $\hat{B}$ are the estimated d.r. directions. Define

$$\text{Accuracy}(\hat{B}, B) = \frac{1}{L} \sum_{i=1}^{L} \|P_B \hat{\beta}_i\|^2 = \frac{1}{L} \sum_{i=1}^{L} \|(BB^T)\hat{\beta}_i\|^2.$$

For example, suppose $B = (\mathbf{e}_1, \mathbf{e}_2)$ where $\mathbf{e_i}$ the i-th coordinator unit vector and $\hat{B} = (\mathbf{e}_2, \frac{1}{\sqrt{2}}\mathbf{e_1} + \frac{\mathbf{1}}{\sqrt{\mathbf{2}}}\mathbf{e_3})$. Then $\text{Accuracy}(\hat{B}, B) = 75\%$ which means that the estimate $\hat{B}$ recovered 75% of the d.r. space B.

Example 1. Consider a binary classification problem in $\mathbb{R}^{10}$ where the d.r. directions are the first two dimensions and the remaining eight dimensions are Gaussian noise. The data in the first two relevant dimensions are plotted in Figure 1(a) for sample size $n = 400$. For this example SIR cannot identify the two d.r. directions because the group averages of the two groups are roughly the same for the first two dimensions, due to the symmetry in the data. Using local averages instead of group average, LSIR can find both directions, see Figure 1(c). So does SAVE and pHd since the high-order moments also behave differently in the two groups.

Next we create a data set for semi-supervised learning by randomly selecting 20 samples, 10 from each group, to be labeled and setting the others to be unlabeled. The directions from PCA where one ignores the labels does not agree with the discriminant directions as shown in Figure 1(b). This implies that the labels need to be taken into account to recover the relevant directions. To illustrate to advantage of a semi-

Method	SAVE	pHd	LSIR ($k = 20$)	LSIR ($k = 40$)
Accuracy	$0.35(\pm 0.20)$	$0.35(\pm 0.20)$	$0.95(\pm .00)$	$0.90(\pm .00)$

Table 1: Average accuracy (and standard error) of various dimension reduction methods for semi-supervised learning in Example 1.

supervised approach we evaluate the accuracy of the semi-supervised version of LSIR with two supervised methods SAVE and pHd which use only the labeled data. In Table 1 we report the average accuracy and standard error of twenty independent draws of the data using the procedure stated above. For one draw of the data we plot the labeled points in Figure 1(a) and the projection onto the top two LSIR directions in Figure 1)(c). These results clearly indicate that LSIR out-performs the other two supervised dimension reduction methods.

Example 2. (Swiss roll) We first generate in $\mathbb{R}^{10}$ the following data sturcture. The first three dimensions are the Swiss roll data (Roweis and Saul, 2000):

$$X_1 = t \cos t, \quad X_2 = 21h, \quad X_3 = t \sin t,$$

where $t = \frac{3\pi}{2}(1 + 2\theta)$, $\theta \sim$ Uniform$(0, 1)$ and $h \sim$ Uniform$([0,\ 1])$. The remaining 7 dimensions of X are independent Gaussian noise. Then each dimension of X is normalized to have unit variance. Consider the following function on this two-dimensional manifold:

$$Y = \sin(5\pi\theta) + h^2 + \epsilon, \tag{5}$$

where we set the noise as $\epsilon \sim \mathrm{N}(0, 0.1^2)$.

We randomly choose n samples as a training set and let n change from 200 to 1000.

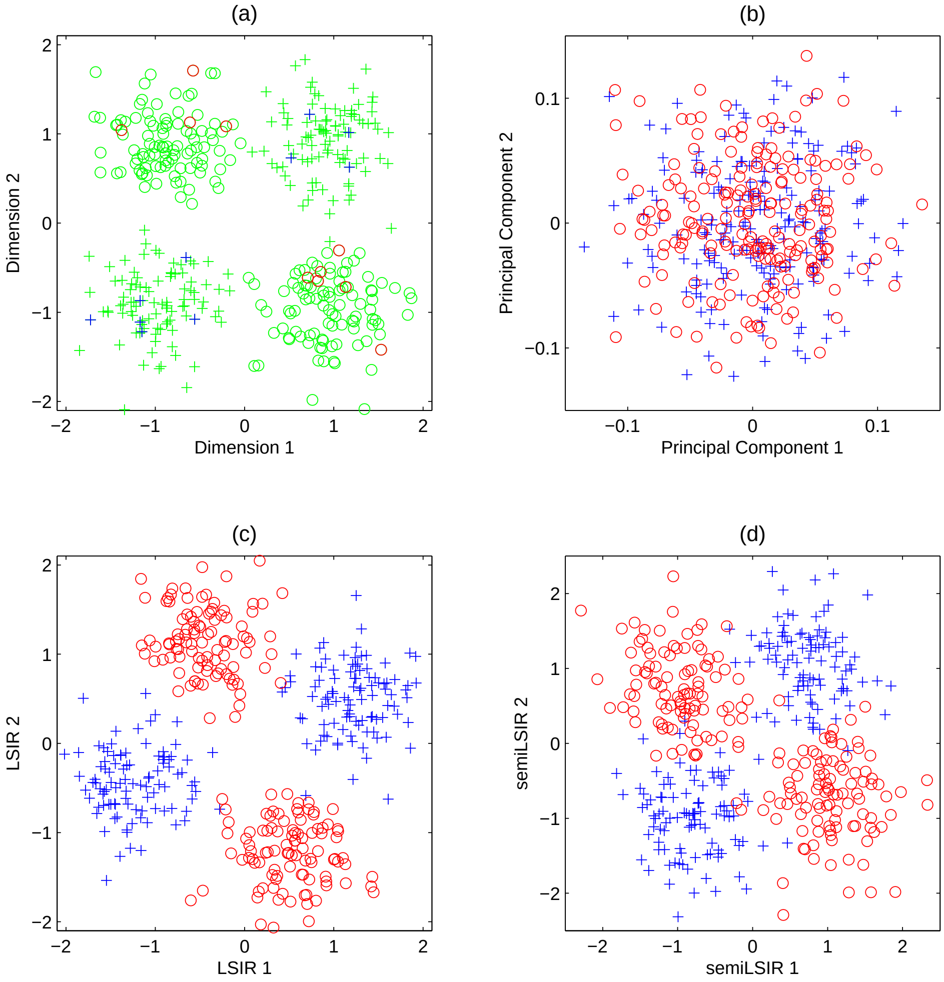

Figure 1: Result for Example 1. (a) Plot of data in the first two dimensions, where ‘+’ corresponds to $y = 1$ while ’o’ corresponds to $y = -1$. The data points in red and blue are labeled and the ones in green are unlabeled in case of the semisupervised setting. (b) Projection of data to the first two PCA directions. (c) Projection of data to the first two LSIR directions when all the $n = 400$ data points are labeled. (d) Projection of the data to the first two LSIR directions when only 20 points as indicated in (a) are labeled.

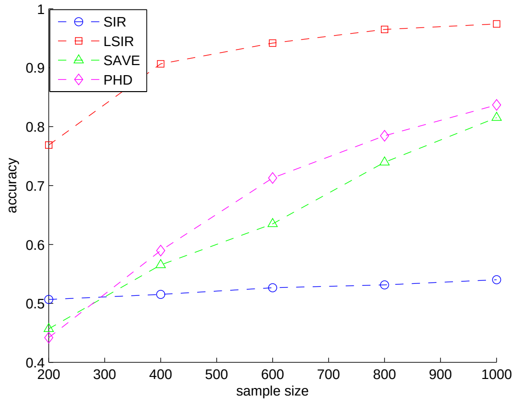

Figure 2: Accuracy on the swiss roll example for a variety of dimension reduction methods.

We compare the performance of LSIR with SIR, SAVE and pHd. For each dimension reduction method, we estimate the d.r. directions and compute the estimation accuracy. The result is showed in Figure 2. SAVE and pHd outperform SIR, but are still much worse compared to LSIR. For LSIR we set the number of slices H and the number of nearest neighbors k as $(5, 5)$ for $n = 200$, $(H, k) = (10, 10)$ for $n = 400$, and $(10, 15)$ for other cases.

Note that the Swiss roll (the first three dimensions) is a benchmark data set in manifold learning, where the goal is to "unroll" the data into the intrinsic two dimensional space. Since LSIR is a linear dimension reduction method we do not expect LSIR to unroll the data, we do expect to retrieve the dimensions relevant to the prediction of Y. With the addition of noisy dimensions manifold learning algorithms will not unroll the data either since the dominant directions are the dimensions corresponding to noise.

Example 3. (Tai Chi) The Tai Chi figure is well known in East Asian culture where the concepts of Yin-Yang provide the intellectual framework for much of ancient Chinese scientific development Ebrey (1993). A 6-dimensional data set for this example is generated as follows: X_1 and X_2 are from the Tai Chi structure as shown in Figure 3(a) where the Yin and Yang regions are assigned class labels $Y = -1$ and $Y = 1$ respectively. $X_3, \ldots, X_6$ are independent random noise generated by $\mathrm{N}(0, 1)$.

The Tai Chi data set was first used as a dimension reduction example in (Li, 2000, Chapter 14). The correct d.r. subspace B here is span$(\mathbf{e}_1, \mathbf{e}_2)$. SIR, SAVE and pHd are known to fail for this example. By taking the local structure into account, LSIR can easily retrieve the relevant directions. Following Li (2000), we generate $n = 1000$ samples as the training data, then run LSIR with $k = 10$ and repeat this 100 times. The average accuracy is 98.6% and the result from one run is shown in Figure 3.

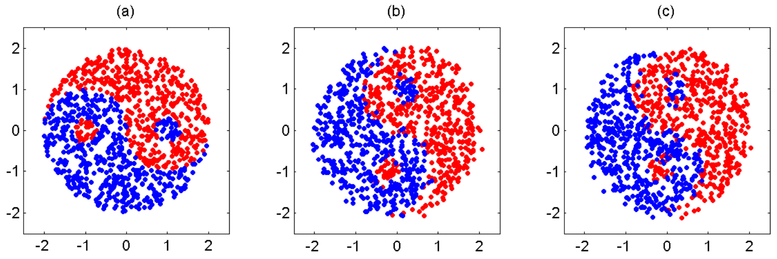

Figure 3: Result for Tai Chi example. (a) The training data in first two dimensions; (b) The training data projected onto the first two LSIR directions; (c) Independent test data projected onto the first two LSIR directions.

For comparison we also applied LFDA to this example. The average estimation accuracy is 82% which is much better than SIR, SAVE and pHd, but still worse than LSIR.

As pointed out by Li (2000), the major difference between Yin and Yang is roughly along the direction $\mathbf{e}_1 + \mathbf{e}_2$, while the difference along the second direction $\mathbf{e}_1 - \mathbf{e}_2$ is subtle. Li (2000) suggested using SIR to identify the first direction and then using SIR-II to identify the second direction by slicing the data based on the value of both y and the projection onto the first direction. LSIR provides a simpler procedure to recover the two directions.

Example 4. Consider a regression model (X, Y) where

$$Y = X_1^3 - ae^{X_1^2/2} + \epsilon$$

with $X_1, \ldots, X_{10}$ iid $\sim$ N$(0, 1)$ and $\epsilon \sim$ N$(0, 0.1^2)$. The d.r. direction is the first coordinate $\mathbf{e}_1$. SIR can easily identify this direction when a is very small. However as a increases, the second term which is a symmetric function becomes dominant and the performance of SIR deteriorates. We will use this example to further study LSIR and compare its behavior with that of SIR.

We first study the effect of the choice of k when a varies. We draw $n = 400$ samples that are split into $h = 10$ slices with each slice containing 40 samples. We measure the error by the angle between the true d.r. direction and the estimate $\hat{\beta}$, which is denoted by $\alpha(\hat{\beta}, \mathbf{e_1})$. The averaged errors from 1000 runs are shown in Figure 4 (a-c) where k ranges from 1 to 40 for $a = 0, 1, 2$, respectively. When $k > 1$, the estimates $\hat{\beta}$ from LSIR are very close to the true d.r. direction. When $a = 0$ which is the case favoring SIR, we can see the errors from LSIR decrease as k increases since LSIR with $k = 40$ is identical to SIR. When $a = 1$, the results from SIR and LSIR agree for a wide range of k. But when $a = 2$, LSIR outperforms SIR.

Next we study how the choice of k influences the estimation of the number L of d.r. directions. In Figure 4 (d-f) we plot the change of the mean of the smallest $(p - L)$ eigenvalues (which theoretically should be 0)

$$\bar{\lambda}_{p-L} = \frac{1}{p-L} \sum_{i=L+1}^{p} \lambda_i$$

with respect to the choice of k as a varies. Recall that in SIR $\bar{\lambda}_{p-L}$ satisfies certain χ^2 distribution and can be used to test the number of d.r. directions (Li, 1991). In LSIR the smallest eigenvalues may not be 0 and λ_{p-L} no longer follows a χ^2 distribution due to localization.

However, we do not consider this as serious drawback. Note that in many applications dimension reduction is used as a preprocessing step. In this case the dimension could be estimated via cross validation. Also, most learning algorithms are sensitive to the accuracy of the d.r. directions but very stable to the addition of 1 or 2 noise directions.

5 Applications to real data

In this section we apply LSIR to several real data sets.

5.1 Digit recognition

The MNIST data set (Y. LeCun, *http://yann.lecun.com/exdb/mnist/*), contains $60,000$ images of handwritten digits $\{0, 1, 2, ..., 9\}$ as training data and $10,000$ images as test data. Each image consists of $p = 28 \times 28 = 784$ gray-scale pixel intensities. This data set is commonly believed to have strong nonlinear structures.

In our simulations, we randomly sampled 1000 images (100 samples for each digit)

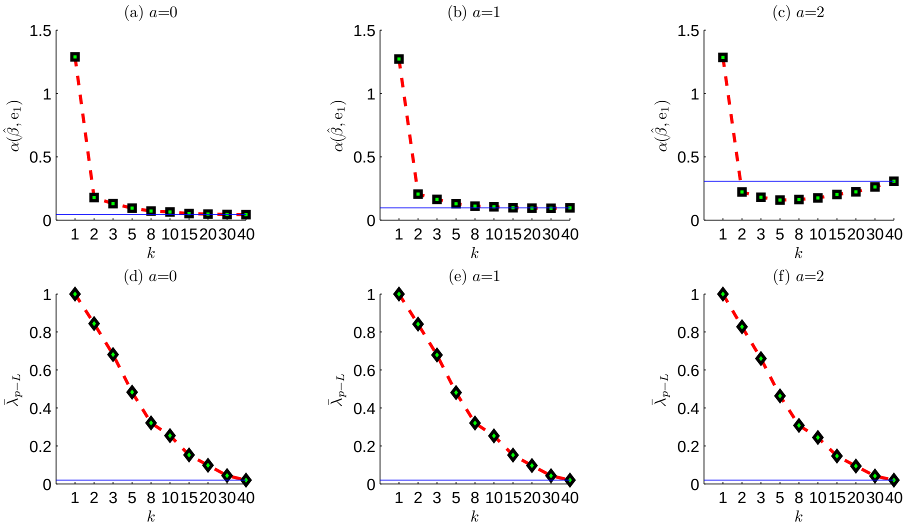

Figure 4: Results for Example 4. The baseline in each plot is for SIR (LSIR with $k = 40$ realizes SIR in this example).

as training data. We applied LSIR and computed $d = 20$ e.d.r. directions. We then projected the training data and 10000 test data onto these directions. Using a k-nearest neighbor classifier with $k = 5$ to classify the test data, we report the classification error over 100 iterations in Table 2. For comparison we report the classification error rate using SIR from Wu et al. (2007). Increasing the number of d.r. directions improves classification accuracy for almost all digits. The improvement for digits $2, \ 3, \ 5, \ 8$ is rather significant. The overall average accuracy for LSIR is comparable with many nonlinear methods.

digit	LSIR	SIR
0	0.04(± 0.01)	0.05 (± 0.01)
1	0.01(± 0.003)	0.03 (± 0.01)
2	0.14(± 0.02)	0.19 (± 0.02)
3	0.11(± 0.01)	0.17 (± 0.03)
4	0.13(± 0.02)	0.13 (± 0.03)
5	0.12(± 0.02)	0.21 (± 0.03)
6	0.04(± 0.01)	0.0816 (± 0.02)
7	0.11(± 0.01)	0.14 (± 0.02)
8	0.14(± 0.02)	0.20 (± 0.03)
9	0.11(± 0.02)	0.15 (± 0.02)
average	0.09	0.14

Table 2: Classification error rate for the digits data for SIR and LSIR.

5.2 Gene expression data

Cancer classification and discovery using gene expression data is an important tool in modern molecular biology and medical science. In these data the expression of thousands of genes are assayed and the number of samples assayed is limited. As is the case of most large p small n problems, dimension reduction can play an essential role in understanding the predictive structure in the data and for inference.

Leukemia classification. In this example we consider the well studied leukemia classification example first developed in Golub et al. (1999). This data consists of 38 training samples and 34 test samples. The training and test samples consist of two types of leukemia, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). An interesting point that we will return to is that the ALL phenotype can be further subdivided into two subsets B-cell ALL samples and T-cell ALL samples. We applied SIR and LSIR to this data. The classification accuracy is similar by predicting the test data with 0 or 1 error. An interesting point is that LSIR automatically realizes subtype discovery while SIR cannot. In Figure 5 we show the projection of the training data onto the first two LSIR directions. One immediately notices that the ALL has two dense clusters implying that ALL has two subtypes. It turns out that the 6-samples cluster are T-cell ALL and the 19-samples cluster is B-cell ALL samples. Note that there are two samples (which are T-cell ALL) cannot be assigned to each subtype only by visualization. This means LSIR only provides useful subclass knowledge for future research but itself may not be a perfect clustering method.

6 Discussion

We incorporated local information into the classical SIR method to develop LSIR. It alleviates many of the degeneracy problems in SIR and increases accuracy, especially when the data has underlying nonlinear structure. When used in classification we see that LSIR can automatically identify subcluster structure. Regularization is added for computational stability and we introduce a semi-supervised version of the method for the addition of unlabeled data and illustrate the utility of the method on utility on synthetic and real data.

LSIR allows us to bridge PCA and SIR by varying the number k of nearest neigh-

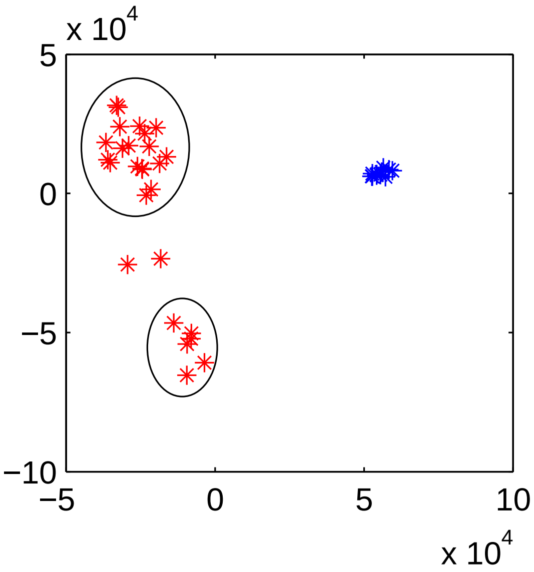

Figure 5: Result for leukemia data using LSIR: training data projected on 2 LSIR directions. Red points are ALL and blue ones are AML

bors varies. The influence of the choice of k is subtle. In cases where SIR works well, k should be chosen to be large so that LSIR performs similar to SIR. Conversely, in case of SIR does not works well for small values of k LSIR outperforms SIR. We find in our simulations a moderate choice of k (between 10 and 20) is is sufficient. A further study of the effect of k and a better theoretical understanding of the method is of interest.

It is straightforward to extend LSIR to kernel models (Wu et al., 2007) to realize nonlinear dimension reduction or reduce the computational complexity in case of $p \gg n$. However, for nonlinear dimension reduction purpose we are skeptical of using a kernel model for LSIR since the LSIR directions already extract local information on a nonlinear manifold.

Appendix. LSIR and PCA

Here we show that the regularized version of LSIR realizes PCA with $k = 1$. Recall that $\hat{\Gamma}_{\text{loc}} = \hat{\Sigma}$ when $k = 1$. The generalized eigen-decomposition problem for LSIR becomes

$$\hat{\Sigma}\beta = \lambda(\hat{\Sigma} + s)\beta. \tag{6}$$

Denote the singular decomposition of $\hat{\Sigma}$ by UDU^T where $D = \text{diag}(d_i)_{i=1}^p$. Then (6) becomes

$$UDU^T\beta = \lambda U(D + s)U^T\beta,$$

which is equivalent to solve

$$D\gamma = \lambda(D + sI)\gamma$$

with $\gamma = U^T\beta$. Since $d/(d + s)$ is increasing with respect to d for any $s > 0$, it is easy to see that the i-th eigenvector β_i is given by the ith column of U which is the ith principal component.

References

Adcock, R. (1878). A problem in least squares. *The Analyst 5*, 53–54.

Belkin, M. and P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation 15*(6), 1373–1396.

Cook, R. and L. Ni (2006). Using intra-slice covariances for improved estimation of the central subspace in regression. *Biometrika 93*(1), 65–74.

Cook, R. and S. Weisberg (1991). Disussion of li (1991). *J. Amer. Statist. Assoc. 86*, 328–332.

Cook, R. and X. Yin (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Aust. N. Z. J. Stat. 43*(2), 147–199.

Donoho, D. and C. Grimes (2003). Hessian eigenmaps: new locally linear embedding techniques for highdimensional data. *PNAS 100*, 5591–5596.

Duan, N. and K. Li (1991). Slicing regression: a link-free regression method. *Ann. Stat. 19*(2), 505–530.

Ebrey, P. (1993). *Chinese Civilization: A sourceboook*. New York: Free Press.

Edegworth, F. (1884). On the reduction of observations. *Philosophical Magazine*, 135–141.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Statistical Society A 222*, 309–368.

Globerson, A. and S. Roweis (2006). Metric learning by collapsing classes. In Y. Weiss, B. Schölkopf, and J. Platt (Eds.), *Advances in Neural Information Processing Systems 18*, pp. 451–458. Cambridge, MA: MIT Press.

Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Hastie, T. and R. Tibshirani (1996). Discrminant adaptive nearest neighbor classification. *IEEE Transacations on Pattern Analysis and Machine Intelligence 18*(6), 607–616.

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

Li, K. C. (1992). On principal hessian directions for data visulization and dimension reduction: another application of stein's lemma. *J. Amer. Statist. Assoc. 87*, 1025–1039.

Li, K. C. (2000). High dimensional data analysis via the sir/phd approach.

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

Sugiyam, M. (2007). Dimension reduction of multimodal labeled data by local fisher discriminatn analysis. *Journal of Machine Learning Research 8*, 1027–1061.

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

Wu, Q., F. Liang, and S. Mukherjee (2007). Regularized sliced inverse regression for kernel models. Technical report, ISDS Discussion Paper, Duke University.

Xia, Y., H. Tong, W. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B 64*(3), 363–410.

Young, G. (1941). Maximum likelihood estimation and factor analysis. *Psychometrika 6*, 49–53.

Zhong, W., P. Zeng, P. Ma, J. S. Liu, and Y. Zhu (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics 21*(22), 4169–4175.