ONLINE LEARNING FOR SUPERVISED DIMENSION REDUCTION

NING ZHANG
Computational Science PhD Program
Middle Tennessee State University
1301 E Main Street, Murfreesboro, TN 37132, USA

QIANG WU*
Department of Mathematical Sciences
Middle Tennessee State University
1301 E Main Street, Murfreesboro, TN 37132, USA

(Communicated by Yunlong Feng)

Abstract. Online learning has attracted great attention due to the increasing demand for systems that have the ability of learning and evolving. When the data to be processed is also high dimensional and dimension reduction is necessary for visualization or prediction enhancement, online dimension reduction will play an essential role. The purpose of this paper is to propose a new online learning approach for supervised dimension reduction. Our algorithm is motivated by adapting the sliced inverse regression (SIR), a pioneer and effective algorithm for supervised dimension reduction, and making it implementable in an incremental manner. The new algorithm, called incremental sliced inverse regression (ISIR), is able to update the subspace of significant factors with intrinsic lower dimensionality fast and efficiently when new observations come in. We also refine the algorithm by using an overlapping technique and develop an incremental overlapping sliced inverse regression (IOSIR) algorithm. We verify the effectiveness and efficiency of both algorithms by simulations and real data applications.

1. Introduction. Dimension reduction aims to explore low dimensional representation for high dimensional data. It helps to promote our understanding of the data structure through visualization and enhance the predictive performance of machine learning algorithms by preventing the “curse of dimensionality”. Therefore, as high dimensional data become ubiquitous in modern sciences, dimension reduction methods are playing more and more important roles in data analysis. Dimension reduction algorithms can be either unsupervised or supervised. Principal component analysis (PCA) might be the most popular unsupervised dimension reduction method. Other unsupervised dimension reduction methods include the kernel PCA, multidimensional scaling, and manifold learning based methods such as isometric mapping and local linear embedding. Unlike the unsupervised dimension reduction, supervised dimension reduction involves a response variable. It finds the intrinsic
lower-dimensional representations that are relevant to the prediction of the response values. Supervised dimension reduction methods can date back to the well known linear discriminant analysis (LDA) while its blossom occurred in the last twenty years. Many approaches have been proposed and successfully applied in various scientific domains; see [25, 10, 26, 44, 37, 15, 28, 42, 43, 11] and the references therein.

We are in a big data era and facing the challenges of big data processing, thanks to the fast development of modern information technology. Among others, two primary challenges are the big volume and fast velocity of the data. When a data set is too big to be stored in a single machine or when the data arrives in real time and information update is needed frequently, analysis of the data in an online manner is necessary and efficient. If the data is simultaneously big and high dimensional, it becomes necessary to develop online learning approaches for dimension reduction. As PCA and LDA are most wildly used in dimension reduction techniques, correspondingly, a bunch of PCA-based and LDA-based online dimension reduction algorithms has been proposed. Incremental PCA have been described in [7, 20, 21, 41, 53, 34, 35]. Incremental LDA have been developed in [33, 46, 38, 17, 52, 24, 23, 39, 45, 40, 8]. In this paper, our purpose is to propose new online learning approaches for supervised dimension reduction. Our motivation is to implement the sliced inverse regression (SIR) in an incremental manner. SIR was proposed in [25] and has become one of the most efficient supervised dimension reduction method. SIR and its refined versions have been found successful in many scientific areas such as bioinformatics, hyperspectral image analysis, and physics; see e.g. [9, 5, 4, 13, 16, 22, 1, 27, 12, 51]. SIR can be implemented by solving an generalized eigen-decomposition problem, \( \Gamma \beta = \lambda \Sigma \beta \), where \( \Gamma \) is a matrix depending on the response variable (whose definition is described in the next section) and \( \Sigma \) is the covariance matrix. To make SIR implementable in an online manner we rewrite it as normal eigen-decomposition problem \( \Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}} \eta = \lambda \eta \) where \( \eta = \Sigma^{-\frac{1}{2}} \beta \) and adopt the ideas from incremental PCA. Some challenges and computational concerns arise in this process. The main contribution of this paper is to overcome these difficulties and design a workable incremental SIR method. We will also refine the method by an overlapping technique and design an incremental overlapping SIR algorithm.

The rest of this paper is arranged as follows. We review SIR algorithm in Section 2 and the incremental PCA algorithm in Section 3. We propose the incremental SIR algorithm and refine it with the overlapping technique in Section 4. Their relations to existing supervised dimension reduction methods are discussed. Simulations are done in Section 5 and we close with discussions in Section 6.

2. Sliced inverse regression. The goal of supervised dimension reduction is to find an intrinsic lower-dimensional subspace that contains all the information to predict the response variable. Assume a multivariate predictor \( x = (x_1, x_2, \ldots, x_p)^\top \in \mathbb{R}^p \) and a scalar response \( y \) are linked by a semi-parametric regression model

\[
y = f(\beta_1^\top x, \beta_2^\top x, \ldots, \beta_K^\top x, \epsilon),
\]

where \( \beta_k \in \mathbb{R}^p \) is a \( p \times 1 \) vector and \( \epsilon \) is the error term independent of \( x \). It implies

\[
y \perp x | \mathbf{B}^\top x,
\]

where \( \perp \) denotes “statistical independence” and \( \mathbf{B} = (\beta_1, \beta_2, \ldots, \beta_K) \) is a \( p \times K \) matrix. The column space of \( \mathbf{B} \) is called the effective dimension reduction (EDR)
space and each $\beta_k$ is an EDR direction. Note $B^\top x$ contains all information for the prediction of $y$. The purpose of supervised dimension reduction is to learn the EDR directions from data.

Unlike the classical regression problem which regresses $y$ against $x$, sliced inverse regression considers regressing $x$ against $y$. With the semi-parametric model (1) and the assumption that $x$ follows an elliptical contour distribution (e.g., normal distribution), it was proved in [25] that the centered regression curve $E[x|y] - E[x]$ falls into the $K$-dimensional subspace spanned by EDR directions $\Sigma \beta_k, k = 1, \ldots, K$. Consequently, all or part of the EDR directions can be recovered by solving a generalized eigen-decomposition problem:

$$\Gamma \beta = \lambda \Sigma \beta,$$

(2)

where

$$\Gamma = E \left[ (E[x|y] - E[x]) (E[x|y] - E[x])^\top \right]$$

is the covariance matrix of the inverse regression curve $E[x|y]$, $\Sigma$ is the covariance matrix of $x$. Each eigenvector associated with a non-zero eigenvalue is an EDR direction.

Given i.i.d observations $\{(x_i, y_i)\}_{i=1}^n$, SIR algorithm can be implemented as follows:

1) Compute sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top.$$

2) Bin the observations into $H$ slices according to $y$ values. For each slice $s_h, h = 1, \ldots, H$, compute the sample probability $\hat{p}_h = \frac{n_h}{n}$ and the sample slice mean $\hat{m}_h = \frac{1}{n_h} \sum_{y_i \in s_h} x_i$. The matrix $\Gamma$ is estimated by

$$\hat{\Gamma} = \sum_{h=1}^H \hat{p}_h (\hat{m}_h - \bar{x}) (\hat{m}_h - \bar{x})^\top.$$

3) Solve the generalized eigen-decomposition problem

$$\hat{\Gamma} \hat{B} = \hat{\Sigma} \hat{B} \Lambda.$$

The EDR directions are estimated by $\hat{\beta}_k, k = 1, 2, \ldots, K$, which are the top $K$ eigenvectors.

This algorithm is not very sensitive to the choice of parameter $H$ provided it is sufficiently larger than $K$ while not greater than $\frac{n}{2}$. Root-$n$ consistency is usually promised. It is suggested samples are evenly distributed into the $H$ slices for best performance.

3. **Incremental PCA.** PCA looks for directions along which the data have largest variances. It is implemented by solving an eigen-decomposition problem

$$\hat{\Sigma} u = \lambda u.$$

(3)

The principal components are the eigenvectors corresponding to largest eigenvalues. Throughout this paper, we assume all eigenvalues are arranged in a descending order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Suppose that we need to retain the top $K$ principal components. Denote $U_K = [u_1, \ldots, u_K], \Lambda_K = \text{diag}(\lambda_1, \ldots, \lambda_K)$. We have a reduced system $\hat{\Sigma} U_K = U_K \Lambda_K$. 
In incremental PCA, after receiving a new coming observation $x_0$, we need to update the reduced eigen-system to a new one

$$\hat{\Sigma}'U'_K \approx U'_K A'_K.$$ 

A “=” is generally impossible unless $\lambda_{K+1} = \ldots = \lambda_p = 0$ in (3).

The idea of updating the system in [20] is as follows. Compute a residual vector

$$v = (x_0 - \bar{x}') - U_K U'_K (x_0 - \bar{x'}),$$

where $\bar{x}'$ is the mean of all observations (including $x_0$). It defines the component of $x_0$ that is perpendicular to the subspace defined by $U_K$. If $x_0$ lies exactly within the current eigenspace, then the residual vector is zero and there is no need to update the system. Otherwise, we normalize $v$ to obtain $\tilde{v} = \frac{v}{\|v\|}$. We may reasonably assume each column vector of $U'_K$ is a linear combination of column vectors of $U_K$ and $\tilde{v}$ – this is exactly true if $\lambda_{K+1} = \ldots = \lambda_p = 0$. This allows us to write

$$[U'_K, u'_{K+1}] = [U_K, \tilde{v}]R.$$

where $R$ is a $(K + 1) \times (K + 1)$ rotation matrix and $u'_{K+1}$ is an approximation of the $(K + 1)$th eigenvector of $\hat{\Sigma}'$. So we have

$$\hat{\Sigma}'[U_K, \tilde{v}]R = [U_K, \tilde{v}]RA'_{K+1},$$

which is equivalent to

$$[U_K, \tilde{v}]^T \hat{\Sigma}'[U_K, \tilde{v}]R = RA'_{K+1}.$$ 

This is an eigen-decomposition problem of dimensionality $K + 1 \ll p$. It solves the rotation matrix $R$ and allows us to update principal components to $U'_K$, given by the first $K$ columns of $[U_K, \tilde{v}]R$. If we need to increase the number of principal components, we can just update the system to $K' = K + 1$ and $U'_{K'} = [U'_K, u'_{K+1}]$. This incremental PCA algorithm was shown convergent to a stable solution when the sample size increases [20].

4. **SIR based online supervised dimension reduction.** In this paper we implement online supervised dimension reduction through incremental implementation of SIR based algorithms. Our idea to develop the incremental sliced inverse regression (ISIR) is motivated by reformulating SIR problem to a PCA problem. To this end, we define $\eta = \Sigma^{-\frac{1}{2}}\beta$, called the standardized EDR direction, and rewrite the generalized eigen-decomposition problem (2) as an eigen-decomposition problem

$$\Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}} \eta = \lambda \eta.$$ 

Note that $\Sigma^{-\frac{1}{2}} \Gamma \Sigma^{-\frac{1}{2}}$ is the covariance matrix of $\Sigma^{-\frac{1}{2}}E[x|y]$. So (5) can be regarded as a PCA problem with data collected for $\Sigma^{-\frac{1}{2}}E[x|y]$. To apply the ideas from IPCA to this transformed PCA problem, however, is not as direct as it looks like. We face two main challenges. First, when a new observation $(x_0, y_0)$ is received, we need to transform it to an observation for the standardized inverse regression curve. This is different from simply standardizing the data. Second, conceptually, we need to update $\Sigma^{-\frac{1}{2}}$ in an online manner in order to standardize the data. Although there did exist some online methods to approximately update $\Sigma^{-\frac{1}{2}}$ in the literature (see for example [18]), they are not exact. There are accuracy and computational concerns if we repeatedly use the estimated $\Sigma^{-\frac{1}{2}}$ to standardize data, compute the intermediate $\eta$, and transform back to the EDR directions. We wish to minimize the use of approximated quantities, avoid the calculation of $\eta$, and update the EDR
space $\mathbf{B}$ directly. This turns out to be possible. In the following, we will describe in detail how we address these challenges and implement ISIR efficiently.

4.1. **Incremental SIR.** Suppose we have $n$ observations in hand with well defined sample slice probabilities $\hat{p}_h$ and means $(\mathbf{m}_h, \bar{y}_h)$ for $h = 1, \ldots, H$, and the eigenvectors $\hat{\mathbf{B}} = [\hat{\beta}_1, \ldots, \hat{\beta}_K]$ of the generalized eigen-decomposition problem $\hat{\mathbf{G}} \beta = \lambda \hat{\Sigma} \beta$. Then with $\Lambda_K = \text{diag}(\lambda_1, \ldots, \lambda_K)$, we have

$$\hat{\mathbf{G}} \hat{\mathbf{B}} = \hat{\Sigma} \hat{\mathbf{B}} \Lambda_K.$$ 

Denote $\Xi = \hat{\Sigma}^{-\frac{1}{2}} \hat{\mathbf{B}}$. We have

$$\hat{\Sigma}^{-\frac{1}{2}} \hat{\mathbf{G}} \hat{\Sigma}^{-\frac{1}{2}} \Xi = \Xi \Lambda_K.$$ 

When we have a new observation $(\mathbf{x}_0, y_0)$, we first locate which slice it belongs to according to the distances from $y_0$ to sample slice mean values $\bar{y}_h$ of the response variable. Let us suppose the distance from $y_0$ to $\bar{y}_k$ is the smallest. So we place the new observation into the slice $k$ and update sample slice probabilities by $\hat{p}_h = \frac{n_{\hat{p}_h}}{n_{\hat{p}_h} + 1}$ for $h \neq k$ and $\hat{p}_k = \frac{n_{\hat{p}_k} + 1}{n_{\hat{p}_k}}$. Let $n_k = n_{\hat{p}_k}$ be the number of observations in slice $k$ before receiving the new observation. For slice mean values we update

$$\mathbf{m}_k' = \frac{n_k}{n_k + 1} \mathbf{m}_k + \frac{1}{n_k + 1} \mathbf{x}_0$$

for slice $k$ only. We can regard $\mathbf{z}_0 = \hat{\Sigma}^{-\frac{1}{2}} \mathbf{m}_k'$ as a new observation for the standardized inverse regression curve $\hat{\Sigma}^{-\frac{1}{2}} \mathbf{E}(\mathbf{x}|y)$. Following the idea of IPCA, we define a residual vector

$$\mathbf{v} = (\mathbf{z}_0 - \hat{\Sigma}^{-\frac{1}{2}} \bar{x}') - \Xi \Xi^\top (\mathbf{z}_0 - \hat{\Sigma}^{-\frac{1}{2}} \bar{x}')$$

and normalize it to $\bar{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$ when $\mathbf{v}$ is not zero. To update the eigen-decomposition system to

$$\hat{\Sigma}'^{-\frac{1}{2}} \hat{\mathbf{G}}' \hat{\Sigma}'^{-\frac{1}{2}} \Xi' = \Xi' \Lambda'_K,$$

we assume $|\Xi', \eta'_{K+1}| = |\Xi, \bar{\mathbf{v}}| \mathbf{R}$ with $\mathbf{R}$ being a $(K+1) \times (K+1)$ rotation matrix and $\eta'_{K+1}$ the $(K+1)$th eigenvector of $\hat{\Sigma}'^{-\frac{1}{2}} \hat{\mathbf{G}}' \hat{\Sigma}'^{-\frac{1}{2}}$. So we have

$$\hat{\Sigma}'^{-\frac{1}{2}} \hat{\mathbf{G}}' \hat{\Sigma}'^{-\frac{1}{2}} [\Xi, \bar{\mathbf{v}}] \mathbf{R} = [\Xi, \bar{\mathbf{v}}] \Lambda'_{K+1},$$

where $\Lambda'_{K+1} = \text{diag}(\Lambda'_K, \lambda'_{K+1})$ and $\lambda'_{K+1}$ is the $(K+1)$th eigenvalue. Multiplying both sides by $[\Xi, \bar{\mathbf{v}}]^\top$, we obtain

$$\left(\hat{\Sigma}'^{-\frac{1}{2}} |[\Xi, \bar{\mathbf{v}}]| \right)^\top \hat{\mathbf{G}}' \left(\hat{\Sigma}'^{-\frac{1}{2}} |[\Xi, \bar{\mathbf{v}}]| \right) \mathbf{R} = \mathbf{R} \Lambda'_{K+1}.$$

Note, however, since $\hat{\Sigma}'^{-\frac{1}{2}}$ cannot be accurately updated, we want to avoid using it. To overcome this challenge, we notice that

$$\hat{\Sigma}' = \frac{n}{n + 1} \hat{\Sigma} + \frac{n}{(n + 1)^2} (\mathbf{x}_0 - \bar{x}) (\mathbf{x}_0 - \bar{x})^\top$$

and the well known Sherman-Morrison formula allows us to update the inverse matrix

$$\hat{\Sigma}^{-1} = \frac{n + 1}{n} \hat{\Sigma}^{-1} - \frac{1}{n} \frac{\hat{\Sigma}^{-1} (\mathbf{x}_0 - \bar{x}) (\mathbf{x}_0 - \bar{x})^\top \hat{\Sigma}^{-1}}{1 + \frac{1}{n+1} (\mathbf{x}_0 - \bar{x})^\top \hat{\Sigma}^{-1} (\mathbf{x}_0 - \bar{x})}.$$
If we store $\hat{\Sigma}^{-1}$ and update it incrementally, we can approximate the quantities in (6) as follows:

$$\hat{\Sigma}'^{-\frac{1}{2}} \Xi \approx \hat{\Sigma}^{-\frac{1}{2}} \Xi = \hat{\cal B},$$

$$\hat{\Sigma}'^{-\frac{1}{2}} \hat{v} \approx \hat{\Sigma}^{-1} (\hat{m}_k' - \bar{x}') = \hat{\cal B} \hat{\cal B}^\top (\hat{m}_k' - \bar{x}'),$$

$$||\hat{v}||^2 = (\hat{m}_k' - \bar{x}')^\top \left(\hat{\Sigma}^{-1} - \hat{\cal B} \hat{\cal B}^\top\right) (\hat{m}_k' - \bar{x}'),$$

$$\bar{v} = \hat{\Sigma}'^{-\frac{1}{2}} \hat{v} = \frac{\hat{\Sigma}'^{-\frac{1}{2}} \hat{v}}{||\hat{v}||} \approx \frac{\left(\hat{\Sigma}^{-1} - \hat{\cal B} \hat{\cal B}^\top\right) (\hat{m}_k' - \bar{x}')}{\sqrt{(\hat{m}_k' - \bar{x}')^\top \left(\hat{\Sigma}^{-1} - \hat{\cal B} \hat{\cal B}^\top\right) (\hat{m}_k' - \bar{x}')}}.$$  

So the problem (6) is approximated by

$$[\hat{\cal B}, \bar{v}]^\top \hat{\Gamma}' [\hat{\cal B}, \bar{v}] = R' R'_{K+1}. \quad (8)$$

Finally notice that the new EDR space $\tilde{\cal B}' = \hat{\Sigma}'^{-\frac{1}{2}} \Xi'$ is the first $K$ columns of $\hat{\Sigma}^{-\frac{1}{2}} [\Xi', \eta'_{K+1}] = \hat{\Sigma}^{-\frac{1}{2}} [\Xi, \bar{v}] R$ and can be approximated by the first $K$ columns of $[\tilde{\cal B}, \bar{v}] R$.

Note that using Sherman-Morrison formula to update $\hat{\Sigma}^{-1}$ is exact and no approximation is involved. We avoided updating the inverse square root of the covariance matrix by using the approximation $\hat{\Sigma}^{-\frac{1}{2}} \approx \hat{\Sigma}'^{-\frac{1}{2}}$. This is the only approximation we used in the algorithm. It can be very accurate when $n$ is large enough because both converge to $\Sigma^{-\frac{1}{2}}$. Therefore, we may expect the convergence of ISIR as a corollary of the convergence of IPCA. However, when $n$ is small, the approximation may be less accurate and result in larger difference between EDR spaces estimated by ISIR and SIR. So we recommend that ISIR be used with a warm start, that is, using SIR first on a small amount of data before using ISIR.

In terms of memory, the primary requirement is the storage of $\hat{\Sigma}^{-1}$, the slice mean matrix $\tilde{\cal M} = [\hat{\bar{m}}_1, \ldots, \hat{\bar{m}}_H]$, and the EDR space $\tilde{\cal B}$. So the memory requirement is $O(p^2 + pH + pK)$. As for the computational complexity, notice that the update of $\hat{\Sigma}^{-1}$ in (7) requires the calculation of $\hat{\Sigma}^{-1} (x_0 - \bar{x})$ and matrix addition and has a complexity of $O(p^2)$. Since we need to store $\tilde{\cal M}$ and update it sequentially, it is not efficient to store and update $\tilde{\Gamma}'$ for either memory or computation consideration. Instead, we use the fact $\tilde{\Gamma}' = \tilde{\cal M}' \cal P' \tilde{\cal M}'^\top$ where $\cal P' = \text{diag}(\tilde{\rho}', \ldots, \tilde{\rho}_H')$ and write

$$[\tilde{\cal B}, \bar{v}]^\top \tilde{\Gamma}' [\tilde{\cal B}, \bar{v}] = [\tilde{\cal B}^\top \tilde{\cal M}', \bar{v}^\top \tilde{\cal M}'] [\tilde{\cal B}^\top \tilde{\cal M}', \bar{v}^\top \tilde{\cal M}']^\top R' R'_{K+1}.$$  

Notice that

$$\bar{v}^\top \tilde{\cal M}' = \frac{(\hat{\Sigma}^{-1} (\hat{m}_k' - \bar{x}'))^\top \tilde{\cal M}' - (\hat{\cal B}^\top (\hat{m}_k' - \bar{x}'))^\top (\hat{\cal B}^\top \tilde{\cal M}')}{\sqrt{(\hat{m}_k' - \bar{x}')^\top (\hat{\Sigma}^{-1} (\hat{m}_k' - \bar{x}')) - (\hat{\cal B}^\top (\hat{m}_k' - \bar{x}'))^\top (\hat{\cal B}^\top (\hat{m}_k' - \bar{x}'))}}.$$
and \( \hat{B}^\top (\hat{m}_k - \bar{x}') \) is just the \( k \)th column of \( \hat{B}^\top \hat{M}' \). The primary computation for the matrix \( [\hat{B}, \hat{v}]^\top \Gamma [\hat{B}, \hat{v}] \) is \( \hat{B}^\top \hat{M}' \) and \( \Sigma^{-1} (\hat{m}_k - \bar{x}') \) which has a complexity of \( O(pKH + p^2) \). The complexity of the eigen-decomposition in (8) is \( O((K+1)^3) \) and to update \( \hat{B} \) to \( \hat{B}' \) requires \( O(p(K+1)^2) \). So the computational complexity for the whole ISIR update is \( O(p^2 + pKH + pK^2 + K^3) \). For a high dimensional problem, this is much smaller than the complexity of \( O(p^3 + p^2n) \) for SIR.

4.2. Refinement by overlapping. In [50] an overlapping technique was introduced to SIR algorithm and shown effectively improving the accuracy of EDR space estimation. It is motivated by placing each observation in two or more adjacent slices to reduce the deviations of the sample slice means \( \bar{\hat{m}}_k \) from the EDR subspace. This is equivalent to using each observation two or more times. In this section, we adopt the overlapping technique to ISIR algorithm above to develop an incremental overlapping sliced inverse regression (IOSIR) algorithm and wish it refines ISIR.

To apply the overlapping idea, we use each observation twice. So when we have \( n \) observations, we duplicate them and assume we have \( N = 2n \) observations. When a new observation \((x_0, y_0)\) is received, we duplicate it and assume we receive two identical observations. Based on the \( y_0 \) value we place the first copy into the slice \( s_k \) if \( \hat{y}_k \) is closest to \( y_0 \) and run ISIR update as described in Section 4.1. Note that if \( \hat{y}_1 < y_0 < \hat{y}_H \), then \( y_0 \) must fall into the interval \([\hat{y}_k, \hat{y}_{k'}] \) with \( k' = k-1 \) or it falls into \([\hat{y}_k, \hat{y}_{k'}] \) with \( k' = k+1 \). So we place the second copy of the new observation to the slice \( s_{k'} \), an adjacent slice to \( s_k \), and run ISIR algorithm again. If \( y_0 \leq \hat{y}_1 \) or \( y_0 > \hat{y}_H \), the second copy will be still placed into \( s_k \) to guarantee all observations are weighted equally.

As OSIR has superior performance over SIR, we expect IOSIR will perform better than ISIR by a price of double calculation time.

We remark that SIR and ISIR can be used for both regression problems and classification problems. But since the concept of “adjacent slice” cannot be defined for categorical values (as is the case in classification problems), IOSIR can only be used for regression problems where the response variable is numeric.

4.3. Relations to existing methods. In the literature online learning approaches for supervised dimension reduction have been studied in the classification setting. They are based on the incremental implementation of LDA and its variants. A review of existing incremental LDA algorithms can be found in [40]. Many of them depends on an incremental QR decomposition; see for example the incremental dimension reduction via QR decomposition (IDR/QR) [46], the incremental LDA using sufficient spanning set (ILDA/SSS) [23], incremental complete LDA (ICLDA) [30]. Some other methods consider the update of a pseudo inverse of the centered data matrix [29] or the matrix \( \Sigma^{-1/2} \) [17].

The relations between SIR and LDA have been well recognized in the literature. Both methods are associated to generalized eigen-decomposition problems. In classification setting two methods are equivalent. At the same time, their difference is also clear: LDA is used for classification problems while SIR focuses on regression problems. In this sense, our ISIR algorithm is novel for regression problems. Moreover, IOSIR can only be used for regression problems and its difference from incremental LDA approaches is even more obvious. In terms of algorithm implementation, our ISIR algorithm did not rely on QR decomposition or the computation...
of $\Sigma^{-1/2}$ for data standardization. This makes our algorithms essentially different from existing ones even when they are applied to classification problems, let alone regression problems.

5. Simulations. In this section, we will verify the effectiveness of ISIR and IOSIR with simulations on artificial and real-world data. Comparisons will be made between them and SIR.

5.1. Artificial data. In the simulations with artificial data, since we know the true model, we measure the performance by the accuracy of the estimated EDR space. We adopt the trace correlation $r(K) = \text{trace}(P_B P_B) / K$ used in [14] as the criterion, where $P_B$ and $P_{\hat{B}}$ are the projection operators onto the true EDR space $B$ and the estimated EDR space $\hat{B}$, respectively.

We consider the following model from [25],

$$y = x_1 (x_1 + x_2 + 1) + \epsilon, \quad (9)$$

where $x = [x_1, x_2, \ldots, x_p]^\top$ follows a multivariate normal distribution, $\epsilon$ follows standard normal distribution and is independent of $x$. It has $K = 2$ effective dimensions with $\beta_1 = (1, 0, 0, \ldots, 0)^\top$ and $\beta_2 = (0, 1, 0, \ldots, 0)^\top$. We conduct the simulation in $p = 10$ dimensional space and select the number of slices as $H = 10$. We give the algorithm a warm start with the initial guess of the EDR space obtained by applying SIR algorithm to a small data set of 40 observations. Then a total of 400 new observations will be fed to update the EDR space one by one. SIR, ISIR, and IOSIR are applied when each observation was fed in and we calculate their trace correlation and cumulative computation time. We repeat this process 100 times. The mean trace correlations of all three methods are reported in Figure 1(a) and the mean cumulative time is in Figure 1(b). We see that ISIR performs quite similar to SIR and IOSIR slightly outperforms both ISIR and SIR. ISIR is much faster than SIR and IOSIR gains higher accuracy by sacrificing on computation time. This verifies the convergence and efficiency of ISIR and IOSIR.

![Figure 1](image-url)
5.2. **Real data application.** We validate the reliability of ISIR on two data sets: Concrete Compressive Strength [47] and Cpusmall (available on https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html).

There have been many proposed algorithms to increase the prediction accuracy on these data sets [47, 48, 49, 32, 19]. We do not intend to outperform those methods. Our goal is to compare the performance of supervised dimension reduction algorithms and verify the effectiveness and correctness of our incremental methods.

The Concrete Compressive Strength data has \( p = 8 \) predictors and 1030 samples. We use \( H = 10 \) and \( K = 3 \) to run SIR, ISIR, and IOSIR. We select 50 observations to warm start ISIR and IOSIR algorithms, then 700 observations are fed sequentially. The left 280 observations are left as test data. After each new observation is received we estimate the EDR space, project the available training set to the estimated EDR space, build a regression model using the k-nearest neighbor method, and compute the MSE on the test data set. This process is repeated 100 times and the average MSE was reported in Figure 2(a). For the Cpusmall data, which has \( p = 12 \) predictors and 8192 samples, we do the experiment with \( H = 10, K = 3 \), 50 observations to warm start ISIR and IOSIR, 2000 observation for sequential training, and 6142 observations for testing. The average MSE was plotted in Figure 2(b). The results indicate both ISIR and IOSIR are as effective as SIR.

6. **Conclusions and discussions.** We proposed two online learning approaches for supervised dimension reduction, namely, ISIR and IOSIR. They are motivated by standardizing the data and reformulating the SIR algorithm to a PCA problem. However, data standardization is only used to motivate the algorithm while not explicitly calculated in the algorithms. We proposed to use Sherman-Morrison formula to online update \( \Sigma^{-1} \) and some approximated calculation to circumvent explicit data standardization. This novel idea played a key role in our algorithm design. Both algorithms are shown effective and efficient. While IOSIR does not
apply to classification problems, it is usually superior over ISIR in regression problems.

We remark that the purpose of ISIR and IOSIR is to keep the dimension reduction accuracy in the situation that a batch learning is not suitable. This is especially the case for streaming data where information update and system involving is necessary whenever new data becomes available. When the whole data is given and one only needs the EDR space from batch learning, ISIR or IOSIR is not necessarily more efficient than SIR because their complexity to run over the whole sample path is $O(p^2N)$, comparable to the complexity $O(p^3 + p^2N)$ of SIR.

There are two open problems worth further investigation. First, the need to store and use $\Sigma^{-1}$ during the updating process is the main bottleneck for ISIR and IOSIR when the dimensionality of the data is ultrahigh. Second, for SIR and other batch dimension reduction methods, many methods have been proposed to determine the intrinsic dimension $K$; see e.g. [25, 36, 6, 2, 3, 31]. They depend on all $p$ eigenvalues of the generalized eigen-decomposition problem and are impractical for incremental learning. We do not have obvious solutions to these problems at this moment and would like to leave them for future research.

REFERENCES

[1] A. Antoniadis, S. Lambert-Lacroix and F. Leblanc, Effective dimension reduction methods for tumor classification using gene expression data, Bioinformatics, 19 (2003), 563–570.
[2] Z. Bai and X. He, A chi-square test for dimensionality with non-Gaussian data, Journal of Multivariate Analysis, 88 (2004), 109–117.
[3] M. P. Barrios and S. Velilla, A bootstrap method for assessing the dimension of a general regression problem, Statistics & Probability Letters, 77 (2007), 247–255.
[4] C. Becker and R. Fried, Sliced inverse regression for high-dimensional time series, in Exploratory Data Analysis in Empirical Research, Springer, (2003), 3–11.
[5] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, European Conference on Computer Vision, (1996), 43–58.
[6] E. Bura and R. D. Cook, Extending sliced inverse regression: the weighted chi-squared test, Journal of the American Statistical Association, 96 (2001), 996–1003.
[7] S. Chandrasekaran, B. S. Manjunath, Y.-F. Wang, J. Winkeler and H. Zhang, An eigenspace update algorithm for image analysis, Graphical Models and Image Processing, 59 (1997), 321–332.
[8] D. Chu, L.-Z. Zhao, M. K.-P. Ng and X. Wang, Incremental linear discriminant analysis: A fast algorithm and comparisons, IEEE Transactions on Neural Networks and Learning Systems, 26 (2015), 2716–2735.
[9] R. D. Cook, Using dimension-reduction subspaces to identify important inputs in models of physical systems, in Proceedings of the section on Physical and Engineering Sciences, American Statistical Association Alexandria, VA, (1994), 18–25.
[10] R. D. Cook and S. Weisberg, Sliced inverse regression for dimension reduction: Comment, Journal of the American Statistical Association, 86 (1991), 328–332.
[11] R. D. Cook and X. Zhang, Fused estimators of the central subspace in sufficient dimension reduction, Journal of the American Statistical Association, 109 (2014), 815–827.
[12] J. J. Dai, L. Lieu and D. Rocke, Dimension reduction for classification with gene expression microarray data, Statistical Applications in Genetics and Molecular Biology, 5 (2006), Art. 6, 21 pp.
[13] L. Elbitski, R. C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M. J. O’Connor, S. Schwartz, W. Miller and F. Chiaromonte, Distinguishing regulatory DNA from neutral sites, Genome Research, 13 (2003), 64–72.
[14] L. Ferré, Determining the dimension in sliced inverse regression and related methods, Journal of the American Statistical Association, 93 (1998), 132–140.
[15] K. Fukumizu, F. R. Bach and M. I. Jordan, Kernel dimensionality reduction for supervised learning, in NIPS, (2003), 81–88.
[16] A. Gannoun, S. Girard, C. Guinot and J. Saracco, Sliced inverse regression in reference curves estimation, Computational Statistics & Data Analysis, 46 (2004), 103–122.

[17] Y. A. Ghassabeh, A. Ghavami and H. A. Moghaddam, A new incremental face recognition system, in IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, (2007), 335–340.

[18] Y. A. Ghassabeh and H. A. Moghaddam, A new incremental optimal feature extraction method for on-line applications, in International Conference Image Analysis and Recognition, Springer, (2007), 399–410.

[19] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman and S. Li, Incremental learning for $\nu$-support vector regression, Neural Networks, 67 (2015), 140–150.

[20] P. M. Hall, D. A. Marshall and R. F. Martin, Incremental eigenanalysis for classification, in BMVC, 98 (1998), 286–295.

[21] P. M. Hall, D. A. Marshall and R. F. Martin, Merging and splitting eigenspace models, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (2000), 1042–1049.

[22] P. He, K.-T. Fang and C.-J. Xu, The classification tree combined with SIR and its applications to classification of mass spectra, Journal of Data Science, 1 (2003), 425–445.

[23] T.-K. Kim, S.-F. Wong, B. Stenger, J. Kittler and R. Cipolla, Incremental linear discriminant analysis using sufficient spanning set approximations, in IEEE Conference on Computer Vision and Pattern Recognition, (2007), 1–8.

[24] M. H. Law and A. K. Jain, Incremental nonlinear dimensionality reduction by manifold learning, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (2006), 377–391.

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

[26] K.-C. Li, On principal hessian directions for data visualization and dimension reduction: Another application of Stein’s lemma, Journal of the American Statistical Association, 87 (1992), 1025–1039.

[27] L. Li and H. Li, Dimension reduction methods for microarrays with application to censored survival data, Bioinformatics, 20 (2004), 3406–3412.

[28] L. Li and X. Yin, Sliced inverse regression with regularizations, Biometrics, 64 (2008), 124–131.

[29] L.-P. Liu, Y. Jiang and Z.-H. Zhou, Least square incremental linear discriminant analysis, in 2009 Ninth IEEE International Conference on Data Mining, (2009), 298–306.

[30] G.-F. Lu, J. Zou and Y. Wang, Incremental learning of complete linear discriminant analysis for face recognition, Knowledge-Based Systems, 31 (2012), 19–27.

[31] G. M. Nkiet, Consistent estimation of the dimensionality in sliced inverse regression, Annals of the Institute of Statistical Mathematics, 60 (2008), 257–271.

[32] A. Öztürk, M. Pala, E. Özbay, E. Kanca, N. Caglar and M. A. Bhatti, Predicting the compressive strength and slump of high strength concrete using neural network, Construction and Building Materials, 20 (2006), 769–775.

[33] S. Pang, S. Ozawa and N. Kasabov, Incremental linear discriminant analysis for classification of data streams, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 35 (2005), 905–914.

[34] C.-X. Ren and D.-Q. Dai, Incremental learning of bidirectional principal components for face recognition, Pattern Recognition, 43 (2010), 318–330.

[35] P. Rodriguez and B. Wohlberg, A matlab implementation of a fast incremental principal component pursuit algorithm for video background modeling, in 2014 IEEE International Conference on Image Processing (ICIP), (2014), 3414–3416.

[36] J. R. Schott, Determining the dimensionality in sliced inverse regression, Journal of the American Statistical Association, 89 (1994), 141–148.

[37] C. M. Setodji and R. D. Cook, K-means inverse regression, Technometrics, 46 (2004), 421–429.

[38] F. X. Song, D. Zhang, Q. Chen and J. Yang, A novel supervised dimensionality reduction algorithm for online image recognition, in Pacific-Rim Symposium on Image and Video Technology, Springer, (2006), 198–207.

[39] J.-G. Wang, E. Sung and W.-Y. Yau, Incremental two-dimensional linear discriminant analysis with applications to face recognition, Journal of Network and Computer Applications, 33 (2010), 314–322.

[40] X. Wang, Incremental and Regularized Linear Discriminant Analysis, Ph.D thesis, National University of Singapore, 2012.
[41] J. Weng, Y. Zhang and W.-S. Hwang, Candid covariance-free incremental principal component analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (2003), 1034–1040.

[42] H.-M. Wu, Kernel sliced inverse regression with applications to classification, *Journal of Computational and Graphical Statistics*, 17 (2008), 590–610.

[43] Q. Wu, S. Mukherjee and F. Liang, Localized sliced inverse regression, *Journal of Computational and Graphical Statistics*, 19 (2010), 843–860.

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

[45] J. Yan, Z. Lei, D. Yi and S. Z. Li, Towards incremental and large scale face recognition, in 2011 International Joint Conference on Biometrics (IJCB), IEEE, (2011), 1–6.

[46] J. Ye, Q. Li, H. Xiong, H. Park, R. Janardan and V. Kumar, IDR/QR: An incremental dimension reduction algorithm via QR decomposition, *KDD ’04 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2004), 364–373.

[47] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, *Cement and Concrete Research*, 28 (1998), 1797–1808.

[48] I.-C. Yeh, Design of high-performance concrete mixture using neural networks and nonlinear programming, *Journal of Computing in Civil Engineering*, 13 (1999), 36–42.

[49] I.-C. Yeh, Modeling slump flow of concrete using second-order regressions and artificial neural networks, *Cement and Concrete Composites*, 29 (2007), 474–480.

[50] N. Zhang, Z. Yu and Q. Wu, Overlapping sliced inverse regression for dimension reduction, preprint, arXiv:1806.08911.

[51] T. Zhang, W. Ye and Y. Shan, Application of sliced inverse regression with fuzzy clustering for thermal error modeling of CNC machine tool, *The International Journal of Advanced Manufacturing Technology*, 85 (2016), 2761–2771.

[52] H. Zhao and P. C. Yuen, Incremental linear discriminant analysis for face recognition, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38 (2008), 210–221.

[53] H. Zhao, P. C. Yuen and J. T. Kwok, A novel incremental principal component analysis and its application for face recognition, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36 (2006), 873–886.

Received for publication March 2019.

E-mail address: ningzhang0123@gmail.com
E-mail address: qwu@mtsu.edu