Computational Statistics and Data Analysis 57 (2013) 223–232

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sparse sufficient dimension reduction using optimal scoring

Tao Wang, Lixing Zhu *

Hong Kong Baptist University, Hong Kong, China
Yunnan University of Finance and Economics, China

ARTICLE INFO

Article history:
Received 21 January 2012
Received in revised form 7 June 2012
Accepted 13 June 2012
Available online 29 June 2012

Keywords:
High dimensionality
Linear discriminant analysis
Optimal scoring
Sliced inverse regression
Sparsity
Sufficient dimension reduction

ABSTRACT

Sufficient dimension reduction is a body of theory and methods for reducing the dimensionality of predictors while preserving information on regressions. In this paper we propose a sparse dimension reduction method to perform interpretable dimension reduction. It is designed for situations in which the number of correlated predictors is very large relative to the sample size. The new procedure is based on the optimal scoring interpretation of the sliced inverse regression method. As a result, the regression framework of optimal scoring facilitates the use of commonly used regularization techniques. Simulation studies demonstrate the effectiveness and efficiency of the proposed approach.

1. Introduction

Dimension reduction is one of the leitmotifs in statistics, enabling classification and regression to be performed in a parsimonious way. Fisher's linear discriminant analysis (LDA) is a popular data-analytic tool for supervised classification and dimension reduction. LDA provides low-dimensional projections of data onto the discriminant directions that capture most of the information about class separation. In the same spirit, dimension reduction without loss of information is one of the dominant themes in the regression setting. Sufficient dimension reduction (SDR; Li, 1991; Cook, 1998) is a methodology for reducing the dimension of predictors while preserving its regression relation with a response. The reduction is also achieved by projecting raw predictors on to a lower-dimensional subspace. In the present paper, we make an attempt to recast a semiparametric SDR approach in the LDA framework such that a simple and easily implemented method can handle sparse dimension reduction. To this end, we describe SDR below.

Let $\mathbf{x}$ be a p-dimensional random vector representing the predictor, and Y be a random variable representing the response. In full generality, SDR seeks to find a set of linear combinations of $\mathbf{x}$, say $\mathbf{B}^T\mathbf{x}$, where $\mathbf{B}$ is a $p \times d$ matrix with $d \leq p$, such that Y depends on $\mathbf{x}$ only through these linear combinations. The subspace span($\mathbf{B}$) is then called a dimension reduction subspace. The intersection of all such subspaces, if also a dimension reduction subspace, is called the central dimension reduction subspace, or the central subspace in short. Under minor conditions (Cook, 1994; Yin et al., 2008), the central subspace exists, and thus, we assume its existence throughout this article. Moreover, the dimension d of the central subspace is treated as known in the subsequent development.

Ever since the introduction of sliced inverse regression (SIR; Li, 1991) and sliced average variance estimation (Cook and Weisberg, 1991), there has been considerable interest in dimension reduction methods (Xia et al., 2002; Ye and Weiss, 2003; Li and Wang, 2007; Cook and Forzani, 2009; Li and Dong, 2009; Zhu et al., 2010; Yin and Li, 2011). SDR has a wide range of applications, and often performs quite well in simple, low-dimensional settings. However, in the high-dimensional setting

* Corresponding author at: Hong Kong Baptist University, Hong Kong, China.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.06.015

where the number of predictors p is large relative to the number of observations n, the SDR methods are not appropriate for two reasons. First, they often require the inverse of the sample covariance matrix that is ill-conditioned or even singular, and thus are not directly applicable. Second, it is difficult to interpret the estimated linear combinations, since they involve all of the predictors.

Attempts have been made to address these problems in the literature. For example, Li et al. (2005) proposed the concept of model-free variable selection based on SDR, and developed some test procedures to assess the contribution of individual predictors. Because their tests are in general incorporated into a stepwise backward or forward search procedure, these methods are computationally intensive and are unsatisfactory in terms of stability. Motivated by the least squares formulation of sliced inverse regression originated by Cook (2004), Ni et al. (2005) combined the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) with sliced inverse regression to produce sparse estimates, while Li and Yin (2008) developed a regularized version that achieves simultaneous predictor selection and dimension reduction and allows $n < p$. Bondell and Li (2009) recently proposed the shrinkage covariance inverse regression estimation. All of them are two-step procedures. First, apply relevant SDR methods to estimate the central subspace, say $\hat{\mathbf{B}}$. Second, determine the shrinkage factor $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p) \in \mathbf{R}^p$ in the span$\{\text{diag}(\boldsymbol{\alpha})\hat{\mathbf{B}}\}$ by a LASSO-type regression. As a result, these methods may fail in high-dimensional situations where the performance of $\hat{\mathbf{B}}$ may not be satisfactory because in the first step, the estimation is based on all the predictors, rather than those relevant ones. Li (2007) studied a sparse SDR by adopting the approach of Zou et al. (2006). Chen et al. (2010) proposed coordinate-independent sparse estimation that can simultaneously achieve sparse sufficient dimension reduction and screen out irrelevant variables efficiently. Both approaches are subspace-oriented because they can incorporate penalization with a broad series of SDR methods. However, they require the inverse of the sample covariance matrix as well, and thus the application is problematic as remarked above.

Cook and Yin (2001) demonstrated that existing dimension-reduction methods in the regression setting can be quite useful for constructing summary plots in the discriminant analysis. For example, SIR can be regarded as a linear discriminant analysis applied to the predictors grouped by the categorical response; see also Pardoe et al. (2007). Note that the linear discriminant analysis is equivalent to a multi-response linear regression using optimal scoring to represent the groups (Hastie et al., 1994). In this paper, we shall show that sliced inverse regression with continuous response variable can be recast into a regression framework via optimal scoring. This provides a way to show the connection between classification and regression in terms of dimension reduction. Motivated by this optimal scoring interpretation for SIR, we then propose a sparse dimension reduction in this paper, which is simple, easy to implement, and applicable to high-dimensional settings. In particular, the new method does not involve the inverse of the covariance matrix of the high-dimensional predictor vector, whereas almost all existing sparse dimension reduction methods involve the inverse matrix such that they have difficulty to handle the "large p, small n" problems.

The paper is organized as follows. We review LDA and introduce optimal scoring for dimension reduction in Section 2.1, and formulate SIR into an optimal scoring problem in Section 2.2. In Section 2.3, the sparse dimension reduction method is proposed and an alternating minimization algorithm is provided. The results of simulation studies are reported in Section 3. Concluding remarks about the proposed method can be found in Section 4.

2. Methodology and main results

2.1. The linear discriminant analysis by optimal scoring

Consider a discrimination problem with G classes and n observations. The training sample consists of measurements $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$ on the p predictors. Their class memberships are known. Let $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ be the $n \times p$ data matrix, and $\mathcal{G}_g$ denote the indices of the observations in the gth class and $n_g = |\mathcal{G}_g|$. Assume that the data matrix $\mathbf{X}$ is centered. The standard estimates of the $p \times p$ within-class and between-class covariance matrices are respectively given by

$$\hat{\boldsymbol{\Sigma}}_W = \frac{1}{n}\sum_{g=1}^{G}\sum_{i\in\mathcal{G}_g}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)^T$$

and

$$\hat{\boldsymbol{\Sigma}}_B = \frac{1}{n}\sum_{g=1}^{G} n_g \hat{\boldsymbol{\mu}}_g \hat{\boldsymbol{\mu}}_g^T,$$

where $\hat{\boldsymbol{\mu}}_g$ is the sample mean vector for class g. Fisher's linear discriminant analysis seeks a low-dimensional projection of the observations such that the between-class variance is large relative to the within-class variance by solving

$$\begin{aligned} &\underset{\boldsymbol{\beta}_k \in \mathbf{R}^p}{\text{maximize}}\ \boldsymbol{\beta}_k^T \hat{\boldsymbol{\Sigma}}_B \boldsymbol{\beta}_k \\ &\text{subject to } \boldsymbol{\beta}_k^T \hat{\boldsymbol{\Sigma}}_W \boldsymbol{\beta}_k = 1, \qquad \boldsymbol{\beta}_k^T \hat{\boldsymbol{\Sigma}}_W \boldsymbol{\beta}_j = 0, \quad j = 1, \ldots, k-1, \end{aligned} \tag{2.1}$$