Journal of Multivariate Analysis 124 (2014) 94–104

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

General directional regression

Zhou Yu [a], Yuexiao Dong [b,*], Mian Huang [c,d]

[a] School of Finance and Statistics, East China Normal University, Shanghai, 200241, China
[b] Department of Statistics, Temple University, Philadelphia, PA, 19122, USA
[c] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China
[d] Key Laboratory of Mathematical Economics (SUFE), Ministry of Education, Shanghai, 200433, China

ARTICLE INFO

Article history:
Received 8 December 2012
Available online 30 October 2013

AMS subject classification:
62H12

Keywords:
General empirical directions
Nonlinear dimension reduction
Permutation test
Sliced inverse regression
Sliced average variance estimation

ABSTRACT

Directional regression is an effective sufficient dimension reduction method which implicitly synthesizes the first two conditional moments. In this paper, we extend directional regression to a general family of estimators via the notion of general empirical directions. Data-driven method is used to identify the optimal estimator within this family. Based on the proposed general directional regression estimators, we develop a new methodology for nonlinear dimension reduction. Improvement of general directional regression over classical directional regression is demonstrated via simulation studies and an empirical study with the wine recognition data.

1. Introduction

Sufficient dimension reduction aims at reducing the dimension of the predictors while keeping all the relevant information for regression or classification. Let $\mathbf{X}$ be a p-dimensional predictor and Y be a 1-dimensional response. Sufficient dimension reduction seeks $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ with $d \leq p$, such that

$$Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\beta}^T \mathbf{X}, \tag{1}$$

that is, Y is independent of $\mathbf{X}$ conditioning on $\boldsymbol{\beta}^T\mathbf{X}$. The column space of $\boldsymbol{\beta}$, or Span($\boldsymbol{\beta}$), is called a dimension reduction space if $\boldsymbol{\beta}$ satisfies (1). Under mild assumptions [1,21] the intersection of all dimension reduction spaces is still a dimension reduction space, which is called the central space and denoted by $\mathcal{S}_{Y|\mathbf{X}}$ [2]. The dimension d of $\mathcal{S}_{Y|\mathbf{X}}$ is known as the structural dimension. Since the seminar paper of sliced inverse regression [9], various methods have been proposed to make inference about the central space, which, among others, include sliced average variance estimation [5], principle Hessian directions [10,3], contour regression [14] and directional regression [13]. Sliced inverse regression and sliced average variance estimation are among the most popular sufficient dimension reduction methods. Without knowing the true link function between the predictor and the response, both methods rely on assumptions about the marginal distribution of the predictor $\mathbf{X}$. For $\boldsymbol{\beta}$ such that Span($\boldsymbol{\beta}$) $= \mathcal{S}_{Y|\mathbf{X}}$, sliced inverse regression requires that

$$E(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}) \text{ is a linear function of } \boldsymbol{\beta}^T\mathbf{X}, \tag{2}$$

which is known as the *linear conditional mean* assumption. Sliced average variance estimation requires the additional *constant conditional variance* assumption

$$\mathrm{Var}(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}) \text{ is a nonrandom matrix.} \tag{3}$$

* Corresponding author.
E-mail address: ydong@temple.edu (Y. Dong).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.10.016

When $\mathbf{X}$ is multivariate normal, both assumptions (2) and (3) are satisfied. For unknown $\boldsymbol{\beta}$ in practice, transformation or reweighting [4] can be applied to the predictor such that the above two assumptions are satisfied for all $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$. Recent work of central solution space [12,7] proposes to model $E(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X})$ as a nonlinear function of $\boldsymbol{\beta}^T\mathbf{X}$ when the assumption (2) is not satisfied.

Sliced inverse regression is known as a first-order method as it involves inverse conditional mean $E(\mathbf{X}|Y)$. Inverse conditional variance $\mathrm{Var}(\mathbf{X}|Y)$, on the other hand, is utilized for the second-order method sliced average variance estimation. Sliced inverse regression and sliced average variance estimation are complement to each other in both the regression and the classification settings. In regression, sliced inverse regression works better when the link function between the response and the predictor is linear or close to linear, while the sliced average variance estimation is more effective with U-shaped link functions. For classification problems, the sliced inverse regression is better at detecting the differences in the group means, while the sliced average variance estimation is more effective in detecting the differences in the group variances. To combine the strength of these methods, both direct and indirect combinations of sliced inverse regression and sliced average variance estimation have been proposed in the literature. Convex combination methods are discussed in [20,24]. To implicitly synthesize sliced inverse regression and sliced average variance estimation, [13] proposed the directional regression, which is shown to enjoy the best overall performance across a wide range of models compared to the convex combination methods. Directional regression is also appealing as it can estimate the central space exhaustively under very mild conditions.

One limitation of directional regression, however, is that it is not guaranteed to improve the sliced inverse regression or sliced average variance estimation. For example, directional regression may be outperformed by sliced inverse regression in the regression setting with a strong linear trend, or in the classification setting with a significant difference in the group means. Similarly, directional regression may be outperformed by sliced average variance estimation when the setting favors the latter. In this paper, we extend directional regression to a family of new sufficient dimension reduction estimators, which include sliced inverse regression and directional regression as special cases. Our estimators naturally synthesize information from both first-order moments and second-order moments. The proposed family of central space estimators inherits the nice theoretical properties of directional regression, such as exhaustiveness and $\sqrt{n}$-consistency.

A natural question is how to identify the optimal estimator or at least an estimator close to the optimal one within this family. For continuous response, we follow the idea of [20], and use bootstrap to choose an optimal estimator within the general directional regression family. For discrete response, we pair each estimator in the proposed family with support vector machine to get a classification rule. The best classification rule, or the rule with the smallest testing error rate, corresponds to our chosen optimal estimator within the family. Our simulation results show that the chosen estimator improves over sliced inverse regression, sliced average variance estimation and directional regression consistently, and performs similarly to the oracle classifier which is based on the true underlying predictor.

Two practical issues about general directional regression are also addressed in this paper. Sequential test is oftentimes used in the dimension reduction literature to determine the unknown structural dimension d. Following [6], we adapt the permutation test to the setting of general directional regression. Classical sufficient dimension reduction aims at finding linear combinations of the original predictors. In the case when we are interested in finding nonlinear features, we use the "kernel trick" and develop the nonlinear version of general direction regression.

The rest of the paper is organized as follows. In Section 2, we introduce the idea of general directional regression via the notion of general empirical directions. We study the sample level estimation and asymptotic properties of the proposed estimators in Section 3. Section 4 discusses the permutation test to determine the unknown structural dimension. Nonlinear dimension reduction via general directional regression is discussed in Section 5. Simulation studies are performed in Section 6, where we use data-driven methods to choose an estimator within the proposed general directional regression family. Section 7 provides an empirical study with the wine recognition data. We conclude this paper with a brief discussion in Section 8. All the technical proofs are given in the Appendix.

2. General directional regression

In this section, we propose a family of general directional regression estimators. It is demonstrated that the proposed estimators can estimate the central space exhaustively under mild assumptions. A reparameterization is also provided for easier implementation in later sections.

2.1. Population level development

Denote $\boldsymbol{\mu} = E(\mathbf{X})$ and $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X})$. Then $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ is the standardized predictor. The relationship between the $\mathbf{Z}$-scale and the $\mathbf{X}$-scale central spaces is captured by $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$, which is known as the invariance law. Let $(\tilde{\mathbf{Z}}, \tilde{Y})$ be an independent copy of $(\mathbf{Z}, Y)$. The main idea behind directional regression is to explore the directional information contained in $\mathbf{A}_0(Y, \tilde{Y}) = E\{(\mathbf{Z} - \tilde{\mathbf{Z}})(\mathbf{Z} - \tilde{\mathbf{Z}})^T|Y, \tilde{Y}\}$. It is shown in [13] that $\mathrm{Span}\{2\mathbf{I}_p - \mathbf{A}_0(Y, \tilde{Y})\} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$.

Without loss of generality, we assume $E(\mathbf{X}) = \mathbf{0}$ hereafter. Given $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, as a random sample of $(\mathbf{X}, Y)$, we define the set of vectors $\{\mathbf{X}_i - c\mathbf{X}_j : 1 \le i < j \le n\}$ to be general empirical directions with $c > 0$. This is a natural extension of the empirical directions introduced by [14], which are the set of vectors $\{\mathbf{X}_i - \mathbf{X}_j : 1 \le i < j \le n\}$. Let $(\tilde{\mathbf{X}}, \tilde{Y})$ be an independent copy of $(\mathbf{X}, Y)$. Our idea is to regress general empirical directions onto $(Y, \tilde{Y})$. For constants $a > 0, b > 0$