Journal of Multivariate Analysis 134 (2015) 99–118

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Robust estimating equation-based sufficient dimension reduction

Jingke Zhou [a], Wangli Xu [a], Lixing Zhu [b,*]

[a] School of Statistics, Renmin University of China, Beijing, China

[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 10 November 2013
Available online 7 November 2014

AMS 2000 subject classifications:
62B05
62G07
62G08
62G35
62G99
62J99

Keywords:
Robust conditional density estimation
Robust nonparametric regression
Robust sufficient dimension reduction
Semiparametric inference
Outliers

ABSTRACT

In this paper, from the estimating equation-based sufficient dimension reduction method in the literature, its robust version is proposed to alleviate the impact from outliers. To achieve this, a robust nonparametric regression estimator is suggested. The estimator is plugged in the estimating equation of the semiparametric sufficient dimension reduction to obtain robust estimator for the central subspace. The asymptotic properties and robustness of the estimator are investigated. Numerical simulation and real data analysis are conducted to examine the performance of the estimators.

1. Introduction

Dimension reduction is becoming more prevalent for high dimensional regression. Sufficient dimension reduction is a promising approach, which captures all the information about response variable in the latent combination of covariates, that is, given the latent variables $B^T X$, we have, $Y \perp\!\!\!\perp X$. There are a number of methods available in the literature. Examples include sliced inverse regression (SIR, [14]), sliced average variance estimation (SAVE, [3]), directional regression (DR, [15]) and discretization–expectation estimation (DEE, [24]), among others.

However, when there are outliers in collected data, having a robust estimation approach against the impact from those data is of importance. Gather et al. [11] first pointed out that SIR is very sensitive to outliers, and at some extreme situation, the SIR estimator gives completely wrong efficient dimension reduction directions simply orthogonal to the true dimension reduction directions. Gather et al. [10] proposed a class of robust SIR procedures, Generalized SIR (GSIR). Among them, Dimension Adjustment MEthod (DAME) is a promising algorithm. Čížek and Härdle [2] studied the robustness of MAVE and OPG that were proposed by Xia et al. [20] and suggested a robust estimator by replacing the least squares loss function by some robust loss function such as the Huber loss function. Zhou [22] proposed a robust canonical correlation estimator,

* Corresponding author.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

http://dx.doi.org/10.1016/j.jmva.2014.10.006

WCANCOR, which downweights outliers by choosing properly weight function such that the influence function of the weighted canonical correlation estimator are bounded.

The aim of this paper is to propose a robust version of the semiparametric sufficient dimension reduction method that was proposed by Ma and Zhu [16,17]. To this end, we first suggest a robust nonparametric regression estimator because in the estimating equation of the semiparametric sufficient dimension reduction method, the traditional nonparametric kernel estimator is used, which are sensitive to outliers (see, e.g. [13]). Inspired by Kim and Scott [13], we investigate a robust Nadaraya–Watson nonparametric regression estimator and construct a robust double kernel local quadratic estimator for the conditional density function and its derivative. Plugging these two estimators into the semiparametric sufficient dimension reduction-based estimating equations to define robust estimator of the central subspace. This is the idea in spirit similar to the robust SIR proposed by Gather et al. [10]. The details will be presented in Sections 3. The estimation consistency is proved. Due to some technical difficulties, the convergence rate and asymptotic normality cannot be achieved in this paper and we leave them for a further study.

This paper is organized as follows. The semiparametric sufficient dimension reduction is briefly reviewed in Section 2. In Section 3, a robust N–W nonparametric regression estimator and a robust conditional density function estimator are proposed. Robust estimators for the central subspace is also defined in this section. In Section 4, several numerical simulations are conducted and a real data is analyzed. All the proofs are given in the Appendix.

2. A brief review of semiparametric sufficient dimension reduction

We give a very brief review on the semiparametric sufficient dimension reduction in this section and see how robust estimation is required against outliers. Ma and Zhu [16] recast the sufficient dimension reduction problem as a semiparametric estimation problem in the following way. For a regression problem, the joint density function of random vector (X, Y) is

$$p(x, y) = p_X(x)p_{Y|X}(x, y) = p_X(x)p_{Y|B^TX}(B^Tx, y)$$

where $p_X(x)$ is the marginal density function of X, and $p_{Y|B^TX}(B^Tx, y)$ is the conditional density function of Y given B^TX. The last equality holds because of the conditional independence $Y \perp\!\!\!\perp X|B^TX$. Then, the distribution of (X, Y) belongs to the class of probability models $\mathcal{P} = \{p(x, y, B) = \eta_1(x) \times \eta_2(B^Tx, y)\}$, where $B \in R^{p\times d}$, $\eta_1(x)$ is any probability density function with respect to x, for any B^Tx, $\eta_2(B^Tx, y)$ is a probability density function respect to y. Thus, sufficient dimension reduction can be viewed as a semiparametric estimation problem where B is the finite dimensional parameter of interest and $\eta = (\eta_1, \eta_2)$ is an infinite dimensional nuisance parameter.

According to the semiparametric inference theory, the estimating equation can be attained by constructing an element belonging to the orthogonal complement space of nuisance tangent space, and thus the generic semiparametric estimating equation for sufficient dimension reduction is derived in [16] as follows:

$$\sum_{i=1}^{n}[g(Y_i, B^TX_i) - E(g(Y, B^TX)|B^TX_i)][\alpha(X_i) - E(\alpha(X)|B^TX_i)] = 0 \tag{1}$$

where $g(Y, B^TX)$ is arbitrary function of (Y, B^TX), and $\alpha(X)$ is arbitrary function of X. Some existing methods such as SIR and DR can be treated as the special cases of the general estimating equation. We write them as Semi-SIR and Semi-DR throughout this paper similar to Ma and Zhu [16]. Specifically, the estimating equation of Semi-SIR is:

$$\frac{1}{n}\sum_{i=1}^{n}[E(X|Y_i) - E(E(X|Y)|B^TX_i)][X_i - E(X|B^TX_i)]^T = 0, \tag{2}$$

and that of Semi-DR is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{3}\{g_j(Y_i, X_i^T\beta) - E[g_j(Y, X^T\beta)|X_i^T\beta]\}\{\alpha_j(X_i) - E[\alpha_j(X)|X_i^T\beta]\} = 0, \tag{3}$$

where

$$\begin{aligned}
g_1(Y, X^T\beta) &= I_p - E(XX^T|Y),\\
g_2(Y, X^T\beta) &= E\{E(X|Y)E(X^T|Y)\}E(X|Y),\\
g_3(Y, X^T\beta) &= E\{E(X^T|Y)E(X|Y)\}E(X|Y),\\
\alpha_1(X) &= -X\{X - E(X|X^T\beta)\}^T,\\
\alpha_2(X) &= \alpha_3(X) = X^T.
\end{aligned}$$

On the other hand, any $p \times d$ matrix B with rank d, can be recomposed as $[B_u^T, B_l^T]^T$, where B_u is a $d \times d$ matrix and B_l is a $(p-d) \times d$ matrix. Assume that B_u is nonsingular without loss of generality. Note that, for any $d \times d$ nonsingular matrix A, the