Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 2621–2635
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On hybrid methods of inverse regression-based algorithms

Li-Xing Zhu[a, b, *], Megu Ohtaki[c], Yingxing Li[d]

[a] *Hong Kong Baptist University, Hong Kong, China*
[b] *Renmin University of China, Beijing, China*
[c] *Research Institute of Radiation Biology, Hiroshima, Japan*
[d] *Cornell University, New York, USA*

Received 10 March 2005; received in revised form 4 August 2005; accepted 20 December 2005
Available online 3 February 2006

Abstract

This paper is two-fold. First, we present a further investigation for the hybrid methods of inverse regression-based algorithms. This investigation provides the evidence of how the hybrids gain the advantages to become more powerful methods than the existing methods when the central dimension reduction (CDR) space is estimated. Second, a Bayes Information Criterion (BIC)-type algorithm is recommended to estimate the dimension of the CDR space. Differing from the popularly used sequential test methods, the new algorithm does not require the asymptotic normality of the estimator of the inverse regression-based matrix. The BIC-based estimator is proven to be consistent. A set of simulations for several typical models were carried out to guide the selection of coefficient in the hybrids.

Keywords: Dimension reduction; Sliced inverse regression; Sliced average variance estimation; Hybrid methods

1. Introduction

Consider regression problem with the response Y and p-dimensional covariable $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$. In sufficient dimension reduction, the main object of interest is the intersection $S_{Y|\boldsymbol{x}}$ of all subspaces $S \subseteq R^p$ such that Y is independent of $\boldsymbol{x}$ when $P_S\boldsymbol{x}$ is given, where $P_S\boldsymbol{x}$ is the projection of $\boldsymbol{x}$ onto the subspace S. $S_{Y|\boldsymbol{x}}$ is called central dimension reduction (CDR) space. An example is the following model proposed by Li (1991):

$$Y = f\left(B^{\mathrm{T}}\boldsymbol{x}, \varepsilon\right), \tag{1.1}$$

where $f(\cdot)$ is an unknown function and ε is the error term independent of $\boldsymbol{x}$, and $B = (\beta_1, \ldots, \beta_k)$ is an unknown $p \times k$ matrix whose columns are of unitary length in L^2-norm and orthogonal to one another. This model includes many popularly used models, for example, the single-index model (Härdle et al., 1993) and the projection pursuit regression model (Friedman and Stuetzle, 1981).

To estimate the CDR space, two inverse regression-based algorithms are popularly used: sliced inverse regression (SIR) (Li, 1991) and sliced average variance estimate (SAVE) (Cook and Weisberg, 1991; Cook, 2000). These

* Corresponding author. Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China. Tel.: +86 3411 7016; fax: +86 3411 5811.
E-mail address: lzhu@hkbu.edu.hk (L.-X. Zhu).

0167-9473/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.csda.2006.01.005

two approaches are, respectively, based on the inverse mean of $\boldsymbol{x}$ given Y and the inverse conditional variance of $\boldsymbol{x}$ given Y. To be precise, denote $\boldsymbol{z} = \Sigma_{\boldsymbol{x}}^{-1/2}(\boldsymbol{x} - E(\boldsymbol{x}))$ and the conditional covariance matrix of $\boldsymbol{z}$, given y by $Cov(\boldsymbol{z}|y) = E\left\{\left[(\boldsymbol{z} - E(\boldsymbol{z}|Y))\,(\boldsymbol{z} - E(\boldsymbol{z}|Y))^{\mathrm{T}}|Y = y\right]\right\}$. Adopting the notion of Cook (1998, pp. 105, 188, 197), we have $S_{Y|\boldsymbol{z}} = \Sigma_{\boldsymbol{x}}^{1/2} S_{Y|\boldsymbol{x}}$. This is the dimension-reduction subspace for the regression of Y on $\boldsymbol{z}$. Furthermore, let $S_{E(\boldsymbol{z}|y)}$ be the inverse regression (IR) subspace spanned by $E(\boldsymbol{z}|y) := E(\boldsymbol{z}|Y = y)$ for all y, and let $S_{I_p - Cov(\boldsymbol{z}|y)}$ be the inverse conditional variance (IRV) subspace spanned by $I_p - Cov(\boldsymbol{z}|y)$ for all y. *SIR* is used to estimate $S_{E(\boldsymbol{z}|y)}$, and *SAVE* is used to estimate $S_{I_p - Cov(\boldsymbol{z}|y)}$. As Li (1991) proved, if the following linear condition is satisfied:

$$E\left(\boldsymbol{z}|P_{S_{Y|\boldsymbol{z}}}\boldsymbol{z}\right) = P_{S_{Y|\boldsymbol{z}}}\boldsymbol{z}, \tag{1.2}$$

where $P_{(\cdot)}$ stands for the projection operator in the standard inner product (see Cook, 1998), then for all y, $E(\boldsymbol{x}|y)$ lies within the space $S_{y|\boldsymbol{z}}$. Hence, $S_{E(\boldsymbol{z}|y)}$ is a subspace of $S_{Y|\boldsymbol{z}}$ and $dim\left(S_{E(\boldsymbol{z}|y)}\right) \leqslant dim\left(S_{Y|\boldsymbol{z}}\right)$, where $dim(\cdot)$ stands for dimension. Similarly, $S_{I_p - Cov(\boldsymbol{z}|y)}$ is also contained in $S_{Y|\boldsymbol{z}}$ (see Cook and Weisberg, 1991) if, together with condition (1.2),

$$Cov\left(\boldsymbol{z}|P_{S_{Y|\boldsymbol{z}}}\boldsymbol{z}\right) = I_p - P_{S_{Y|\boldsymbol{z}}}. \tag{1.3}$$

Both *SIR* and *SAVE* have pros and cons. *SIR* is relatively insensitive to the choice of slice number. Li (1991) demonstrated this through the numerical study. Zhu and Ng (1995) proved that under some regularity conditions, the asymptotic normality of *SIR* holds in a large range of number of slices, that is, from $\mathrm{O}\left(\sqrt{n}\right)$ to $n/2$. Zhu and Fang (1996) studied kernel estimation, in which the bandwidth selection is also flexible in a fairly large range. When the regression is an odd function, *SIR* works well, as indicated by the simulation in Li (1991). However, when the regression function is symmetric or the CDR space relates to the variance of error, *SIR* does not work well. A typical example is $y = \left(\beta_1^{\mathrm{T}}\boldsymbol{z}\right)^2 + \left(\beta_2^{\mathrm{T}}\boldsymbol{z}\right)^2\varepsilon$, where $\boldsymbol{z}$ and ε have the respective normal distributions $\mathrm{N}(0, I_p)$ and $\mathrm{N}(0, 1)$, $\beta_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$ and $\beta_2 = (0, 1, \ldots, 0)^{\mathrm{T}}$, that is, $y = z_1^2 + z_2^2\varepsilon$. We can easily prove that $E(\boldsymbol{z}|y) \equiv 0$ through $E\left(\boldsymbol{z}|z_1^2, z_2^2, \varepsilon\right) \equiv 0$ and $E(\boldsymbol{z}|y) = E\left(E\left(\boldsymbol{z}|z_1^2, z_2^2, \varepsilon\right)|y\right)$, therefore, $dim(S_{E(\boldsymbol{z}|y)}) = 0$. This issue was also investigated by Li (1991, Remark 4.5) and the discussants of Li's paper. If $y = f_1\left(\beta_1^{\mathrm{T}}\boldsymbol{z}\right) + f_2\left(\beta_2^{\mathrm{T}}\boldsymbol{z}\right)\varepsilon$, and $f_i(\cdot)$ are even functions with respect to $\beta_i^{\mathrm{T}}\boldsymbol{z}$, then $dim\left(S_{E(\boldsymbol{z}|y)}\right)$ may be zero. As to *SAVE*, it can handle the above model well.The IRV space may not be a null space as has been discussed by Li (1991, Remark 4.5 and Rejoinder) and the discussants of Li's paper (see, e.g. Cook and Weisberg, 1991). Cook (2000) investigated *SAVE* in more detail. Furthermore, Cook and Critchley (2000) and Ye and Weiss (2003) showed that under the regularity conditions IR space is a subspace of *IRV* space. Therefore, theoretically, under regularity conditions, *SAVE* should be a more powerful method to estimate CDR space than *SIR*. However, *SAVE* is less robust because it uses higher moments, and is sensitive to the choice of the number of slices. Cook (2000) heuristically stated that, like a tuning parameter in non-parametric estimation, the number of data per slice should be large enough to allow a reasonable estimation of the intra-slice conditional covariance matrices of $\boldsymbol{z}$ given y. Our empirical studies also indicate this (see Section 5 below).

Hence, how to gain the advantages from these two algorithms is of interest. When investigating the relationship among *SIR*, SIR_{II} and *SAVE* in the rejoinder to the discussions on his paper, Li (1991) suggested a hybrid algorithm in an ad hoc manner, that is, a hybrid algorithm of SIR^2 and SIR_{II}, that is, $SIRII_{a_1} = (1 - a_1)\,SIR^2 + a_1 SIR_{II}$ for $0 \leqslant a_1 \leqslant 1$. In this hybrid, *SIR* is indirectly used. Gannoun and Saracco (2003a) studied the asymptotic properties of $SIRII_{a_1}$ and Gannoun and Saracco (2003b) considered a single-index model when $SIRII_{a_1}$ is used. Ye and Weiss (2003) considered the hybrid method of *SIR* and pHd, and briefly discussed the hybrid of *SIR* and *SAVE*. That is, the algorithms $(1 - a)SIR + apHd$, and $(1 - a)SIR + aSAVE$ are used to estimate CDR space. Although *SAVE* is more comprehensive than *SIR*, Cook and Critchley (2000) and Ye and Weiss (2003) pointed out that the increased flexibility comes with a price that relatively straightforward structure that is manifest through the mean $E(\boldsymbol{z}|y)$ is harder to detect with *SAVE* than with *SIR*. Therefore, properly selecting the coefficient may make a more powerful algorithm to estimate CDR space.

Hence with the hybrid idea, there are two issues: how to properly select the coefficient, and how to estimate the dimension of CDR space. Ye and Weiss (2003) proposed a bootstrap selection for the coefficient when the hybrid of *SIR* and pHd is used, and estimation for the dimension. However, when slicing estimation is used, the consistency of the bootstrap estimators is still an open problem. This is because the concomitants $\boldsymbol{z}_i$ associated with the ordered y_i are