ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Fused sliced average variance estimation

Hyoin An, Sungmin Won, Jae Keun Yoo *

Department of Statistics, Ewha Womans University, Seoul 03760, Republic of Korea

ARTICLE INFO

Article history:
Received 9 March 2017
Accepted 30 June 2017
Available online xxxx

AMS 2000 subject classifications:
62G08
62H05

Keywords:
Fusing
Inverse regression
Sliced average variance estimation
Sufficient dimension reduction

ABSTRACT

In this paper, we propose an approach to combine the kernel matrices constructed by sliced average variance estimation (SAVE) with various numbers of slices. The proposed approach is called fused sliced average variance estimation (FSAVE). By fusing the information by usual SAVE applications with different slice numbers, the sensitivity to slices can be reduced, so the structural dimension estimation can be improved. Numerical studies confirm this, and a real data analysis is presented.

1. Introduction

The main goal of sufficient dimension reduction (SDR) is to replace the original predictor $\mathbf{X}$ with a lower-dimensional linearly transformed predictor $\mathbf{M}^{\mathrm{T}}\mathbf{X}$ without losing any information about the distribution of $Y|\mathbf{X}$. This is equivalently expressed as $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{M}^{\mathrm{T}}\mathbf{X}$, where $\perp\!\!\!\perp$ indicates stochastic independence. The central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is the intersection of all subspaces $\mathcal{S}(\mathbf{M})$ satisfying $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{M}^{\mathrm{T}}\mathbf{X}$, where $\mathcal{S}(\mathbf{A})$ stands for a subspace spanned by the columns of $\mathbf{A} \in \mathbb{R}^{p\times q}$. Then SDR pursues the estimation of $\mathcal{S}_{Y|\mathbf{X}}$.

Two methods of sliced inverse regression (SIR; Li, 1991) and sliced average variance estimation (SAVE; Cook and Weisberg, 1991) are classical but most popular among other SDR methods. We will discuss SAVE in detail in later section. To implement SIR and SAVE in practice, a categorization of Y, called *slicing*, is crucial. Critical drawback of the slicing is that there are no optimal numbers which can be used as a thumb rule. Many studies show that improper numbers of the slices mislead the results. This problem is more severe in SAVE than in SIR, because the former utilizes the information of the second moment of $\mathbf{X}|Y$, but the latter does the first moment.

Cook and Zhang (2014) proposed a simple solution to this problem with the application to SIR. They combine sample kernel matrices of SIR from many different numbers of slices. According to Cook and Zhang (2014), the combining procedure to SIR results in equally good basis estimates for the different numbers of slices, varying from 3 to 15.

It should be noted that the Cook–Zhang approach is applicable to SAVE. The primary interest of the paper is to develop a fused approach for SAVE. It is not guaranteed that the fused SAVE always yields robust basis estimates to various numbers of slices like SIR. However, more information on $\mathcal{S}_{Y|\mathbf{X}}$ can be obtained by combining the sample kernel matrices by SAVE, so it can be expected to provide robust basis estimates for mild numbers of slices, say 3 to 6 and to have better asymptotic behaviors in the dimension estimation of $\mathcal{S}_{Y|\mathbf{X}}$ than usual SAVE application.

* Corresponding author.
E-mail address: peter.yoo@ewha.ac.kr (J.K. Yoo).

http://dx.doi.org/10.1016/j.jkss.2017.06.003
1226-3192/

Please cite this article in press as: An, H., et al., Fused sliced average variance estimation. Journal of the Korean Statistical Society (2017), http://dx.doi.org/10.1016/j.jkss.2017.06.003.

The organization of the paper is as follows. In Section 2, the method of SAVE and required conditions are briefly discussed. A fused approach of SAVE is proposed in Section 3. Numerical studies and a real data example are presented in Section 4. We summarize our work in Section 5.

2. Sliced average variance estimation

To explain SAVE, the original predictor $\mathbf{X}$ is standardized as follows: $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - E(\mathbf{X}))$, where $\Sigma = \mathrm{cov}(\mathbf{X})$. Define $\mathcal{S}_{Y|\mathbf{Z}}$ as the central subspace for a regression of $Y|\mathbf{Z}$. Denote d as the true structured dimension of $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{Z}}$. Let $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_z$ be $p \times d$ orthonormal basis matrix for $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{Z}}$, respectively. Then, we have $\mathcal{S}_{Y|\mathbf{X}} = \Sigma^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$, equivalently, $\boldsymbol{\eta} = \Sigma^{-1/2}\boldsymbol{\eta}_z$. For rigorous proofs of the relationship between $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{Z}}$, one can read Cook (1998, proposition 6.3).

Consider the following two conditions called linearity and constant variance condition, respectively:

A1. $E(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{\boldsymbol{\eta}_z}}\mathbf{Z})$ is linear in $\mathbf{P}_{\boldsymbol{\eta}_z}\mathbf{Z}$.

A2. $\mathrm{cov}(\mathbf{Z}|\mathbf{P}_{\boldsymbol{\eta}_z}\mathbf{Z})$ is constant.

If $\mathbf{X}$ has an elliptically contoured distribution, condition A1 is satisfied and A2 approximately holds. If $\mathbf{X}$ follows a multivariate normal distribution, A1 and A2 are guaranteed to hold. In the case that the two conditions fail, $\mathbf{X}$ can often be one-to-one transformed to induce these conditions.

Then conditions A1 and A2 force the following relationship to hold for $\mathrm{cov}(\mathbf{Z}|Y)$:

$$\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y) = \mathbf{P}_{\boldsymbol{\eta}_z}\{\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y)\}\mathbf{P}_{\boldsymbol{\eta}_z} \in \mathcal{S}_{Y|\mathbf{Z}}. \tag{1}$$

Rigorous discussion on Eq. (1) is given in Cook and Weisberg (1991). Therefore, a proper subset of $\mathcal{S}_{Y|\mathbf{Z}}$ can be produced by varying y in $\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y = y)$:

$$\mathcal{S}\{\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y)\} = \mathcal{S}[E\{\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y)\}^2] \subseteq \mathcal{S}_{Y|\mathbf{Z}}.$$

An approach to estimate $\mathcal{S}_{Y|\mathbf{Z}}$ through $\mathbf{M}_{\mathrm{SAVE}} = E\{\mathbf{I}_p - \mathrm{cov}(\mathbf{Z}|Y)\}^2$ is called sliced average variance estimation (SAVE; Cook and Weisberg, 1991). According to Ye and Weiss (2003, section 2.2) SAVE provides a more comprehensive estimation of $\mathcal{S}_{Y|\mathbf{X}}$ than SIR, which utilizes the information of $E(\mathbf{Z}|Y)$, in sense that a subspace constructed by the former contains that by the latter. The estimation of $\mathrm{cov}(\mathbf{Z}|Y)$ by using the exact distribution of $\mathbf{Z}|Y$ is not realistic in practice, not only because the regression is for studying $Y|\mathbf{Z}$, not $\mathbf{Z}|Y$, but also because a strong parametric assumption for $\mathbf{Z}|Y$ is required. Note that $\mathrm{cov}(\mathbf{Z}|Y)$ can be nonparametrically replaced by usual sample moment estimators of the covariance of $\mathbf{Z}$ within each category of Y, if Y is categorical. Following this, the estimation of $\mathrm{cov}(\mathbf{Z}|Y)$ would be simple through a categorization of Y, called *slicing*, if Y is many-valued or continuous. First obtain $\tilde{Y}$ by slicing Y with h levels. Hereafter, h stands for the number of slices in the SAVE application. Then compute the sample inverse covariance within each slice s, $s = 1, \ldots, h$:

$$\hat{\mathrm{cov}}(\mathbf{Z}|\tilde{Y} = s) = \frac{1}{n_s}\sum_{\tilde{Y}_i = s}(\hat{\mathbf{Z}}_{i\in s} - \bar{\mathbf{Z}}_s)(\hat{\mathbf{Z}}_{i\in s} - \bar{\mathbf{Z}}_s)^{\mathrm{T}},$$

where n_s is the sample size of the category s and $\bar{\mathbf{Z}}_s = \frac{1}{n_s}\sum_{\tilde{Y}=s}\hat{\mathbf{Z}}_i$.

Then construct a sample kernel matrix $\hat{\mathbf{M}}_{\mathrm{SAVE}}$:

$$\hat{\mathbf{M}}_{\mathrm{SAVE}} = \sum_{s=1}^{h}\frac{n_s}{n}\left\{\mathbf{I}_p - \hat{\mathrm{cov}}(\mathbf{Z}|\tilde{Y} = s)\right\}\left\{\mathbf{I}_p - \hat{\mathrm{cov}}(\mathbf{Z}|\tilde{Y} = s)\right\}.$$

Then, by spectral decomposition of $\hat{\mathbf{M}}_{\mathrm{SAVE}}$, the eigenvectors corresponding to non-zero eigenvalues form an orthonormal basis of $\mathcal{S}_{Y|\mathbf{Z}}$.

3. Fused approach of SAVE

In this section, we propose a fused approach of SAVE. We start the section with defining that

$$\mathbf{M}^{(h)}_{\mathrm{FSAVE}} = (\mathbf{M}^{(2)}_{\mathrm{SAVE}}, \ldots, \mathbf{M}^{(h)}_{\mathrm{SAVE}}), \tag{2}$$

where $\mathbf{M}^{(h)}_{\mathrm{SAVE}}$ stands for the kernel matrix of SAVE with h slices. The case of $h = 1$ is naturally excluded, because it simply returns null matrix. Also the case of $h = 2$ will not be considered, because $\mathbf{M}^{(2)}_{\mathrm{FSAVE}} = \mathbf{M}^{(2)}_{\mathrm{SAVE}}$.

Theoretically, we have $\mathcal{S}(\mathbf{M}^{(k)}_{\mathrm{SAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$ for $k = 2, 3, \ldots, h$, which directly implies that

$$\mathcal{S}(\mathbf{M}^{(k)}_{\mathrm{SAVE}}) \subseteq \mathcal{S}(\mathbf{M}^{(h)}_{\mathrm{FSAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}, \quad k = 2, \ldots, h.$$

Therefore, $\mathbf{M}^{(h)}_{\mathrm{FSAVE}}$ is a possible kernel matrix to estimate $\mathcal{S}_{Y|\mathbf{Z}}$, so we have

$$\Sigma^{-1/2}\mathcal{S}(\mathbf{M}^{(h)}_{\mathrm{FSAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{X}}.$$

Please cite this article in press as: An, H., et al., Fused sliced average variance estimation. Journal of the Korean Statistical Society (2017), http://dx.doi.org/10.1016/j.jkss.2017.06.003.