Journal of Statistical Planning and Inference 139 (2009) 952 – 961

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Partial central subspace and sliced average variance estimation

Yongwu Shao*, R. Dennis Cook, Sanford Weisberg

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

ARTICLE INFO

Article history:
Received 16 May 2007
Received in revised form
29 April 2008
Accepted 6 June 2008
Available online 11 June 2008

Keywords:
Partial IRE
Marginal tests

ABSTRACT

Sliced average variance estimation is one of many methods for estimating the central subspace. It was shown to be more comprehensive than sliced inverse regression in the sense that it consistently estimates the central subspace under mild conditions while slice inverse regression may estimate only a proper subset of the central subspace. In this paper we extend this method to regressions with qualitative predictors. We also provide tests of dimension and a marginal coordinate hypothesis test. We apply the method to a data set concerning lakes infested by Eurasian Watermilfoil, and compare this new method to the partial inverse regression estimator.

Published by Elsevier B.V

1. Introduction

Consider a regression problem where the response Y is univariate and the predictors consist of an $\mathbb{R}^p$-valued continuous random vector X and a qualitative random variable W. Sufficient dimension reduction here seeks to replace the predictor X with $\eta^{\mathrm{T}}X$ without loss of information on Y, where η is a $p \times d$ matrix with $d \leqslant p$. The column space span(η) of η is called a partial dimension reduction subspace (Chiaromonte et al., 2002) if all the information about Y in X is contained in $\eta^{\mathrm{T}}X$, or equivalently if it satisfies

$$Y \perp\!\!\!\perp X | (\eta^{\mathrm{T}}X, W), \tag{1}$$

where $\perp\!\!\!\perp$ indicates independence. Chiaromonte et al. (2002) showed that when the marginal support of the probability density function of X is path-connected, the intersection of all partial dimension reduction subspaces is itself a partial dimension reduction subspace and it is called the partial central subspace. The concept of a partial central subspace is an extension of the concept of a central subspace (Cook, 1994), which is defined similarly except that we do not have a discrete predictor W and thus (1) becomes $Y \perp\!\!\!\perp X | \eta^{\mathrm{T}}X$. We denote the partial central subspace and the central subspace by $\mathscr{S}_{Y|X}^{(W)}$ and $\mathscr{S}_{Y|X}$, respectively.

Some early methods proposed to estimate the central subspace are sliced inverse regression (SIR) proposed by Li (1991), sliced average variance estimation (SAVE) proposed by Cook and Weisberg (1991) and inverse regression estimation (IRE) proposed by Cook and Ni (2005). IRE estimates the same subspace as SIR, but it makes use of a nonlinear least squares objective function and it is optimal in a class of inverse regressions. All of these methods look at the inverse regression $X|Y$ to learn about the central subspace, with SIR and IRE using only first moments $\mathrm{E}(X|Y)$ while SAVE uses second moments and implicitly uses first moments as well. SAVE has been shown to be more comprehensive than SIR in the sense that it consistently estimates $\mathscr{S}_{Y|X}$ under mild conditions while SIR may estimate only a proper subset of $\mathscr{S}_{Y|X}$ (Cook and Lee, 1999). Since IRE and SIR estimate the same subspace, SAVE is also more comprehensive than IRE.

* Corresponding author at: School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA. Tel.: +1 612 625 9003; fax: +1 612 624 8828.
E-mail addresses: shaoy@imsweb.com, ywshao@stat.univ.edu (Y. Shao).

0378-3758/$ - see front matter Published by Elsevier B.V
doi:10.1016/j.jspi.2008.06.002

Most methods proposed to estimate the partial central subspace are based on methods that estimate the central subspace. For example, Chiaromonte et al. (2002) provided a specific methodology called partial SIR, which is based on SIR, to estimate the partial central subspace. Based on IRE, Wen and Cook (2007) proposed a method called partial IRE, which is optimal in a class of inverse regressions. Ni and Cook (2006) extended the applicability of partial SIR by removing the restrictive homogeneous covariance condition.

In this paper we propose partial SAVE, which is an extension of SAVE to estimate the partial central subspace. Partial SAVE is more comprehensive that partial SIR and partial IRE, in the same way that SAVE is more comprehensive than SIR and IRE. Thus, partial SAVE should be a useful new method for dimension reduction in regressions with qualitative predictors. We review SAVE in Section 2. The algorithm for partial SAVE is presented in Section 3. In Section 4 we develop an asymptotic test for dimension under partial SAVE. Simulation results are presented in Section 5 and an illustrative data analysis is given in Section 6. We conclude with a discussion of related issues in Section 7.

2. Review of sliced average variance estimation

In this section we briefly review Cook and Weisberg's development of SAVE. Suppose η is a basis of the central subspace $\mathscr{S}_{Y|X}=\mathrm{span}(\eta)$. Assume $\Sigma=\mathrm{cov}(X)$ is nonsingular, and let $Z=\Sigma^{-1/2}[X-\mathrm{E}(X)]$ be the predictor standardized to have zero mean and identity covariance.

Sliced average variance estimation is defined by looking at the conditional distribution of $Z|Y$, and in particular by examining the conditional variance $\mathrm{cov}(Z|Y)$. Assume that Y is discrete and it can take values of $1,\dots,s$ where $s\geqslant 2$. If Y is continuous we can always slice it into s non-overlapping slices to get a discrete version. Define

$$M=\mathrm{E}[I_p-\mathrm{cov}(Z|Y)]^2. \tag{2}$$

Shao et al. (2007) showed that $\mathrm{span}(M)=\mathscr{S}_{Y|Z}$ under the following three assumptions:

(A) *Linearity condition*: $\mathrm{E}(X|\eta^{\mathrm{T}}X)$ is a linear function of $\eta^{\mathrm{T}}X$.
(B) *Constant variance condition*: $\mathrm{cov}(X|\eta^{\mathrm{T}}X)$ is a non-random matrix.
(C) *Coverage condition*: For any nonzero $\beta\in\mathscr{S}_{Y|X}$, either $\mathrm{E}(\beta^{\mathrm{T}}X|Y)$ or $\mathrm{cov}(\beta^{\mathrm{T}}X|Y)$ is nondegenerate.

The coverage condition (C) was suggested by a referee and replaces a somewhat stronger condition we originally used that requires the existence of fourth moments. Condition (C) is identical to the condition required by directional regression for dimension reduction (Li and Wang, 2007).

In the sample we replace the x_i by $z_i=\widehat{\Sigma}^{-1/2}(x_i-\bar{x})$, where $\bar{x}$ is the average of the x_i and $\widehat{\Sigma}=n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}}$. We can estimate the matrix M in (2) consistently as

$$\widehat{M}=\sum_{k=1}^{s}\widehat{f}_k[I_p-\widehat{\mathrm{cov}}(Z|Y=k)]^2.$$

Here and throughout, $\widehat{\mathrm{cov}}$ indicates the usual sample covariance matrix operator, so $\widehat{\mathrm{cov}}(Z|Y=k)$ is the sample estimate of $\mathrm{cov}(Z|Y)$ in slice k based on n_k observations in that slice, and $\widehat{f}_k=n_k/n$ is the sample fraction of observations in slice k. Let $\widehat{\beta}_j$ denote the eigenvector corresponding to the jth largest eigenvalue of $\widehat{M}$. Let $d=\dim(\mathscr{S}_{Y|Z})$. If d is known we can use the span of $\widehat{\Sigma}^{-1/2}(\widehat{\beta}_1,\dots,\widehat{\beta}_d)$ to estimate $\mathscr{S}_{Y|X}$ consistently. In practice we can use a sequential test to estimate d.

Suppose that $\mathscr{V}$ is a user-selected subspace of $\mathbb{R}^p$. The marginal coordinate test (Cook, 2004) is $\mathscr{S}_{Y|X}\subseteq\mathscr{V}$ versus $\mathscr{S}_{Y|X}\nsubseteq\mathscr{V}$, where $\dim(\mathscr{V})=m<p$. For example, suppose we have three predictors (X_1,X_2,X_3) and we contemplate removing X_3 from our regression by testing the hypothesis $Y\perp\!\!\!\perp X_3|(X_1,X_2)$. In this case $\mathscr{V}=\mathrm{span}((1,0,0)^{\mathrm{T}},(0,1,0)^{\mathrm{T}})$ and our goal is to test $\mathscr{S}_{Y|X}\subseteq\mathscr{V}$. Let α_x be a user-selected basis for $\mathscr{V}$ expressed in a $p\times m$ matrix of full rank m. Define $\widehat{\alpha}=\widehat{\Sigma}^{1/2}\alpha_x(\alpha_x^{\mathrm{T}}\widehat{\Sigma}\alpha_x)^{-1/2}$, and let $\widehat{H}$ be orthonormal basis for the orthogonal complement of $\mathrm{span}(\widehat{\alpha})$, expressed as a $p\times(p-m)$ matrix. Shao et al. (2007) constructed the following test statistic:

$$T_n(\widehat{H})=\frac{n}{2}\sum_{k=1}^{s}\widehat{f}_k\,\mathrm{tr}[(\widehat{H}^{\mathrm{T}}(I_p-\widehat{\mathrm{cov}}(Z|Y=k))\widehat{H})^2].$$

They showed that $T_n(\widehat{H})$ converges in distribution to $\chi^2[(p-m)(p-m+1)/2]$ when X is normal, and converges to a weighted χ^2 in general, where the weights are given by the eigenvalues of $\mathrm{cov}((\alpha_x^{\perp})^{\mathrm{T}}X\otimes(\alpha_x^{\perp})^{\mathrm{T}}X)$, and $\alpha_x^{\perp}$ is an orthonormal basis of the orthogonal complement of α_x under the inner product Σ.

In addition Shao et al. (2007) also considered the marginal dimension hypothesis, which tests $\dim\mathscr{S}_{Y|X}=m$ against $\dim\mathscr{S}_{Y|X}>m$, without specifying which m dimensional subspace of $\mathbb{R}^p$ is intended. For the marginal dimension hypothesis $d=m$, define the test statistic $\widehat{\theta}=(\widehat{\beta}_{d+1},\dots,\widehat{\beta}_p)$, then $T_n(\widehat{\theta})$ also converges to $\chi^2[(p-m)(p-m+1)/2]$ when X is normal, and converges to a weighted χ^2 in general, where the weights are given the eigenvalues of $\mathrm{cov}(\theta^{\mathrm{T}}Z\otimes\theta^{\mathrm{T}}Z)$, and θ is an orthonormal basis of the orthogonal complement of $\mathscr{S}_{Y|Z}$.

3. Extension to partial sliced average variance estimation

The extension of SAVE to partial SAVE allows inclusion of qualitative predictors into a problem. Suppose W is a qualitative predictor with K levels represented by $\{1, \dots, K\}$. If more than one qualitative predictor is available, we combine them into one predictor consisting of all K possible crossed levels. A subspace span(η) is called a partial dimension reduction subspace if it satisfies $Y \perp\!\!\!\perp X | (\eta^{\mathrm{T}} X, W)$. When the intersection of all partial dimension reduction subspaces is still a partial dimension subspace we call it a partial central subspace (Chiaromonte et al., 2002). Throughout this paper we assume the partial central subspace exists and we denote it by $\mathscr{S}^{(W)}_{Y|X}$.

We shall follow Chiaromonte et al. (2002) and Ni and Cook (2006) and use (X_w, Y_w) to indicate a generic pair distributed like $(X, Y)|(W = w)$. By this notation, $S_{Y_w|X_w}$ is the central subspace given $W = w$ and

$$Z_w = \Sigma_w^{-1/2}[X_w - \mathrm{E}(X_w)],$$

where $\Sigma_w = \mathrm{cov}(X_w) > 0$. Chiaromonte et al. (2002) proved the following identity:

$$\mathscr{S}^{(W)}_{Y|X} = \bigoplus_{w=1}^{K} \mathscr{S}_{Y_w|X_w},$$

where the operator $\oplus$ on two subspaces S_1 and S_2 is defined as $S_1 \oplus S_2 = \{v_1 + v_2 : v_1 \in S_1, v_2 \in S_2\}$.

Under the condition that the subpopulation covariance matrices are constant, $\Sigma_w = \Sigma_{\mathrm{pool}}$ for all w, they showed that

$$\mathscr{S}^{(W)}_{Y|X} = \Sigma_{\mathrm{pool}}^{-1/2} \bigoplus_{w=1}^{K} \mathscr{S}_{Y_w|Z_w}.$$

Assume that X_w satisfies the linearity condition and constant variance condition, then under the coverage condition $\mathscr{S}_{Y_w|Z_w}$ is equal to the column space of $\mathrm{E}[I_p - \mathrm{cov}(Z|Y, W = w)]^2$, $\mathscr{S}^{(W)}_{Y|X}$ is equal to the column space of

$$M^{(W)} = \mathrm{E}[I_p - \mathrm{cov}(Z|Y, W)]^2,$$

therefore we only need to construct an estimate of $M^{(W)}$ to estimate $\mathscr{S}^{(W)}_{Y|X}$.

Suppose that we have a random sample of size n for (X, Y, W) from the total population, among which there are n_w observations in subpopulation w, and n_{wk} points have $Y_w = k$. Let $p_w = \mathrm{pr}(W = w)$, $\widehat{p}_w = n_w/n$, $f_{wk} = \mathrm{pr}(Y_w = k)$ and $\widehat{f}_{wk} = n_{wk}/n_w$. Let h_w denote the number of slices in subpopulation w. Following the notations by Ni and Cook (2006), let x_{wki} be the ith observation on X in slice k of subpopulation w, let $\bar{x}_{w\bullet\bullet}$ be the average in subpopulation w,

$$\bar{x}_{w\bullet\bullet} = \frac{1}{n_w} \sum_{k=1}^{h_w} \sum_{i=1}^{n_{wk}} x_{wki},$$

and let $\bar{x}_{wk\bullet}$ be the average of n_{wk} points in slice k of subpopulation w. Let $\widehat{\mathrm{cov}}(X|W = w)$ be the sample variance in subpopulation w. Then we can construct the following estimate of the common covariance:

$$\widehat{\Sigma}_{\mathrm{pool}} = \sum_{w=1}^{K} \frac{n_w}{n} \widehat{\mathrm{cov}}(X|W = w).$$

In the sample we replace the x_{wki}'s with $z_{wki} = [\widehat{\mathrm{cov}}(X|W = w)]^{-1/2}(x_{wki} - \bar{x}_{w\bullet\bullet})$, and let $\widehat{\mathrm{cov}}(Z|W = w, Y = k)$ be the sample covariance of z_{wki}, $i = 1, \dots, n_{wk}$. Then $M^{(W)}$ can be estimated by

$$\widehat{M}^{(W)} = \frac{1}{n} \sum_{w=1}^{K} \sum_{k=1}^{h_w} n_{wk}[I_p - \widehat{\mathrm{cov}}(Z|W = w, Y = k)]^2.$$

Let $\widehat{\beta}_j$ denote the eigenvector corresponding to the jth largest eigenvalue of $\widehat{M}^{(W)}$. Then the span of $\widehat{\Sigma}_{\mathrm{pool}}^{-1/2}(\widehat{\beta}_1, \dots, \widehat{\beta}_d)$ is a consistent estimator of $\mathscr{S}^{(W)}_{Y|X}$.

4. Tests and distribution

4.1. Marginal dimension test

Since d is generally unknown, a method is required for estimating it. The typical procedure is based on tests of the marginal dimension hypothesis $d = m$ versus $d > m$. Starting with $m = 0$, test $d = m$. If the hypothesis is rejected, increment m by one and test again, stopping with the first nonsignificant result. The corresponding value of m is the estimate $\widehat{d}$ of d.

The test statistic we are going to construct is as follows. Let $\widehat{\theta} = (\widehat{\beta}_{m+1}, \dots, \widehat{\beta}_p)$. Define

$$T(\widehat{\theta}) = \frac{1}{2} \sum_{w=1}^{K} \sum_{k=1}^{h_w} n_{wy} \operatorname{tr}[(\widehat{\theta}^{\mathrm{T}} (I_p - \widehat{\operatorname{cov}}(Z|W=w, Y=k))\widehat{\theta})^2]. \tag{3}$$

The next theorem concerns the asymptotic distribution of $T(d)$ and is proved in detail in Appendix A. When $K=1$, we have only one group, and the result reduces to Theorem 2 in Shao et al. (2007).

Theorem 1. *Assume that X_w's are normally distributed with the same covariance. Let $d = \operatorname{rank}(M^{(W)})$. Assume that $Y_w \perp\!\!\!\perp Z_w | M^{(W)} Z_w$. Then the asymptotic distribution of $T(\widehat{\theta})$ defined in* (3) *is $\chi^2[(h-K)(p-m)(p-m+1)]$.*

4.2. Marginal coordinate test

As in Shao et al. (2007), we can construct the marginal coordinate hypothesis test for partial SAVE. Suppose we want to test $\mathcal{S}^{(W)}_{Y|X} \subseteq \mathcal{V}$ versus $\mathcal{S}^{(W)}_{Y|X} \not\subseteq \mathcal{V}$, where $\dim(\mathcal{V}) = m < p$. Let α_x be a user-selected basis for $\mathcal{V}$ expressed in a $p \times m$ matrix of full rank m. Define $\widehat{\alpha} = \widehat{\Sigma}^{1/2}_{\text{pool}} \alpha_x (\alpha_x^{\mathrm{T}} \widehat{\Sigma}_{\text{pool}} \alpha_x)^{-1/2}$, and let $\widehat{H}$ be orthonormal basis for the orthogonal complement of $\operatorname{span}(\widehat{\alpha})$, expressed as a $p \times (p-m)$ matrix. Define the following test statistic:

$$T_n(\widehat{H}) = \frac{1}{2} \sum_{w=1}^{K} \sum_{k=1}^{h_w} n_{wy} \operatorname{tr}[(\widehat{\alpha}^{\mathrm{T}} (I_p - \widehat{\operatorname{cov}}(Z|W=w, Y=k))\widehat{\alpha})^2]. \tag{4}$$

For this test statistic, we have the following theorem, which reduces to Theorem 3 in Shao et al. (2007) when $K=1$.

Theorem 2. *Assume that X_w's are normally distributed with the same covariance. Let $d = \operatorname{rank}(M^{(W)})$. Assume that for each group, the coverage condition holds and $\mathcal{S}_{Y_w|X_w} \subseteq \mathcal{V}$. Then the asymptotic distribution of $T(\widehat{H})$ defined in* (4) *is $\chi^2[(h-K)(p-m)(p-m+1)]$.*

We can also derive the general test when X's are not normal. We did not include it here because the conditions required for the test are quite complicated.

5. Simulations

In this section, we present results from a simulation study to investigate properties of partial IRE and partial SAVE. We do not include partial SIR here because it is dominated by partial IRE (Wen and Cook, 2007). Results are based on three models in which the predictor $X \in \mathbb{R}^p$ is sampled from one of two normal populations that are indicated by W, $W=-1$ or 1, $X \sim N(0, I_p)$:

$$Y = WX_1 + \varepsilon, \tag{5}$$

$$Y = W(X_1^2 + X_2) + \varepsilon, \tag{6}$$

$$Y = W(X_1 + \mu)^2 + \varepsilon, \tag{7}$$

where in each model $\varepsilon \sim N(0, 0.1^2)$, and μ is a constant. The first two models are used for evaluating the marginal tests, and the third is used for investigating the estimation accuracy. The first and the third are one-dimensional models and the second is a two-dimensional model. In each sampling configuration half of the sample was generated from each subpopulation. The simulation was repeated 1000 times for partial SAVE and 100 times for partial IRE, which requires substantially more computations.

5.1. Level

Table 1(a) contains results from marginal dimensional and coordinate test for models (5) and (6). In the marginal coordinate test, we test the null hypothesis $\mathcal{S}^{(W)}_{Y|X} \subseteq \mathcal{V}$ versus the alternative hypothesis $\mathcal{S}^{(W)}_{Y|X} \not\subseteq \mathcal{V}$, where $\mathcal{V} = \operatorname{span}(e_1, e_2)$ and e_i is the canonical basis vector with a one in the ith place and zero elsewhere. The null hypothesis is true in both models. From the results we can see that the actual level is very close to the theoretical level.

In the marginal dimension test, we test the null hypothesis $d=1$ versus $d>1$ for model (5) and $d=2$ versus $d>2$ for model (6). In both tests the null hypothesis is true. The coverage condition for partial SAVE requires that either $E(X_1|Y) \neq 0$ or $\operatorname{cov}(X_1|Y)$ be non-constant, therefore it is satisfied in both models. The coverage condition for partial IRE requires that $E(X_1|Y) \neq 0$, which is violated in model (6). As a result, the test based on partial IRE in model (6) performs poorly, which is reflected in the simulations. The test-based partial SAVE does not share this problem.

Table 1
Estimated levels and powers in simulated data sets.

n	Model (5)				Model (6)			
	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4
(a) *Estimated levels as percentages of nominal 5% tests*								
100	3.9	48	3	13	4.2	39	0.3	5
200	5.3	19	5.4	12	5	15	1.8	0
400	4.3	16	4.6	7	4.4	16	3.4	0
800	5.3	8	4.7	5	4.6	8	4	0
(b) *Estimated powers as percentages of nominal 5% tests*								
100	97.5	100	83	100	99.9	100	14	21
200	100	100	100	100	100	100	66.5	13
300	100	100	100	100	100	100	95.7	6
400	100	100	100	100	100	100	99.9	4

T_1, marginal coordinate test based on partial SAVE; T_2, marginal coordinate test based on partial IRE; T_3, marginal dimension test based on partial SAVE; T_4, marginal dimension test based on partial IRE.

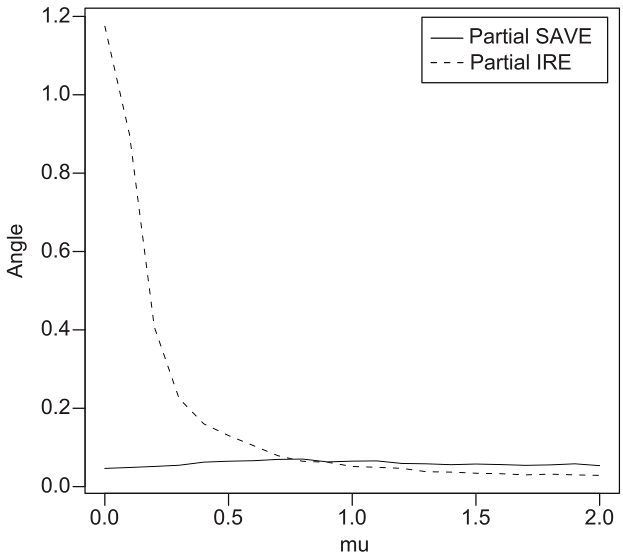

Fig. 1. Comparison of partial SAVE and partial IRE for different μ in model (7). The vertical axis is the angle in radians between the estimated direction and the true direction.

5.2. Power

Table 1(b) contains estimated powers of the marginal coordinate and dimension tests. In the marginal coordinate test, we test the null hypothesis $\mathcal{S}_{Y|X}^{(W)} \subseteq \mathcal{V}$ versus the alternative hypothesis $\mathcal{S}_{Y|X}^{(W)} \not\subseteq \mathcal{V}$, where $\mathcal{V} = \text{span}(e_2)$. The null hypothesis is false in both models. From the results we can see that the actual power of the test based on partial SAVE performs well.

In the marginal dimension test, we test the null hypothesis $d=0$ versus $d>0$ for model (5) and $d=1$ versus $d>1$ for model (6). In both tests the null hypothesis is true. We expect that both partial SAVE and partial IRE will perform reasonably well in model (5) because all required conditions are satisfied. However, partial IRE should have a high power because it requires estimating few objects. For model (6), the test based on partial SAVE performs well for large n, while the test based on partial IRE does not perform well.

5.3. Estimation accuracy

For model (7) we measured the accuracy of estimation through the average values over 200 replications of the angles in radians between the estimated direction vector and the true direction vector. The angle of two vectors is defined as the arc cosine of the inner product of the two vectors normalized to have length one. If the angle is close to zero, this means that the estimation is good. Again we compared the proposed method with the partial IRE. The value of μ varied from zero to two. The results are plotted in Fig. 1. When μ is small, partial SAVE outperforms partial IRE. This is because when $\mu=0$ the coverage condition for partial IRE fails and as a consequence it can estimate only a proper subset of the partial central subspace. Partial SAVE is more comprehensive than partial IRE and does not share this problem. This reflects the primary motivation for the partial SAVE methodology. When μ gets larger, partial IRE begins to catch up and overtake partial SAVE. In this case the coverage conditions hold for both partial SAVE and partial IRE. The estimation accuracy of partial SAVE suffers a bit since it needs to estimate more parameters than partial IRE.

6. Eurasian watermilfoil data

Eurasian watermilfoil is an exotic feathery submerged aquatic plant that can quickly form thick mats in shallow areas of lakes in North America. These mats can interfere with swimming and entangle propellers, which hinders boating, fishing, and waterfowl hunting (Smith and Barko, 1990). To study and control this nuisance species, the Aquatic Plant Management program in the Minnesota Department of Natural Resources (MNDNR) collected a sample of 1866 lakes in Minnesota. The response variable Milfoil is binary and it indicates whether the lake has been infested by watermilfoil. We have seven continuous predictors and one categorical predictor, as described in Table 2.

The categorical predictor Lakeclass is a classification of the lakes, and it is numbered in order of increasing Morphoedaphic indices, or MEI (Schupp, 1992). The lakes in northeastern Minnesota have higher class numbers than the lakes in southwestern Minnesota. Since our methodology needs to estimate a few covariance matrices in each group, we need a relatively large number of observations in each group when the number of predictors is large. Generally we need a larger sample size than model-based discriminant analysis because we do not assume a traditional model. Since there were too few lakes for each class, and we need a relative large sample size for each group, we reclassify the lakes into two categories according to the class number. The lakes with class number between 20 and 32 were classed to one group and the rest were classed to another group. By doing this each group has roughly 900 observations. Since our methodology requires the predictors be normally distributed, we need to transform some of the predictors. We used the multivariate Box–Cox procedure (see Cook and Weisberg, 1999, Section 13.2), and the powers needed for the transformations are listed in Table 2.

Partial IRE is not useful for this problem because the response is binary, and it can find at most one direction in each group. We applied the partial SAVE methodology to the Eurasian watermilfoil data with transformed predictors. The marginal dimension test results are shown in Table 3. We can see that, at level 5%, the dimension of the partial central subspace is inferred to be 2. Therefore $d = 2$ linear combinations of the predictors are needed to characterize the regression.

The first two directions estimated by partial SAVE are $(0.088, -0.128, 0.076, 0.003, 0.293, 0.015, 0.079)$ and $(-0.022, 0.099, -0.117, 0.010, 0.138, 1.481, 1.131)$. If we multiply the coefficients by the standard deviation of the corresponding predictors, we can get the relative scales of the coefficients related to the explanatory variables. The relative scales of the coefficients in the first directions are $(1.02, -0.17, 0.06, 0.07, 0.05, 0.00, 0.07)$. Therefore the first direction is mainly dominated by the first variable, that is, the cube root of distance. The relative scales of the coefficients in the second direction are $(-0.25, 0.13, -0.09, 0.26, 0.03, 0.43, 1.04)$. The major component in the second direction is the cube root of TALK, followed by the cube root of Secchi.

Figs. 2 and 3 show plots of the two estimated directions computed by partial SAVE. From the plots we can see that the infested and uninfested lakes have different means (1.55 and 2.92, respectively) in the first direction and have different variances (0.33 and 1.06, respectively) in the second direction.

Table 2
Variable list in the Eurasian watermilfoil data.

Variable	Variable description	Transformation
Distance	Distance to the nearest lake with Eurasian watermilfoil	1/3
Acres	Area of the lake, in acres	log
Maxd	The maximum depth of the lake in meters	log
PctLitt	The percentage of the lake that is in the littoral zone	1
SDF	Shoreline development factor	−1
Secchi	Secchi depth in meters	1/3
TALK	Total alkalinity in mg/L	1/3
Lakeclass	A categorical variable which takes values from 20 to 45	1

Note: The transformation is based on the multivariate Box–Cox procedure (see Cook and Weisberg, 1999, Section 13.2).

Table 3
Significance levels for marginal dimension test results for the Eurasian watermilfoil data.

m	$T_n(\widehat{\theta})$	df	p-Value
0	123.17	56	0.000
1	69.47	42	0.005
2	35.55	30	0.223
3	20.37	20	0.435

df: degree of freedoms.

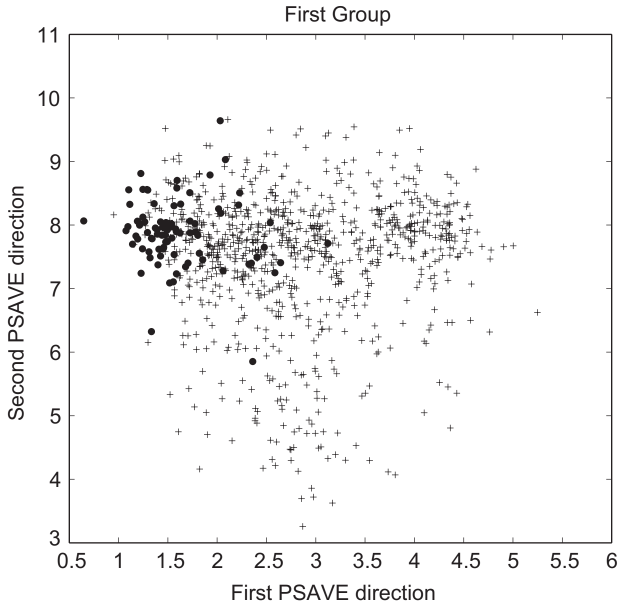

Fig. 2. Plot of the first two directions of the Eurasian watermilfoil data, where "•" denotes that the lake has been infested by water milfoil, "+" denotes that the lake has not been infested.

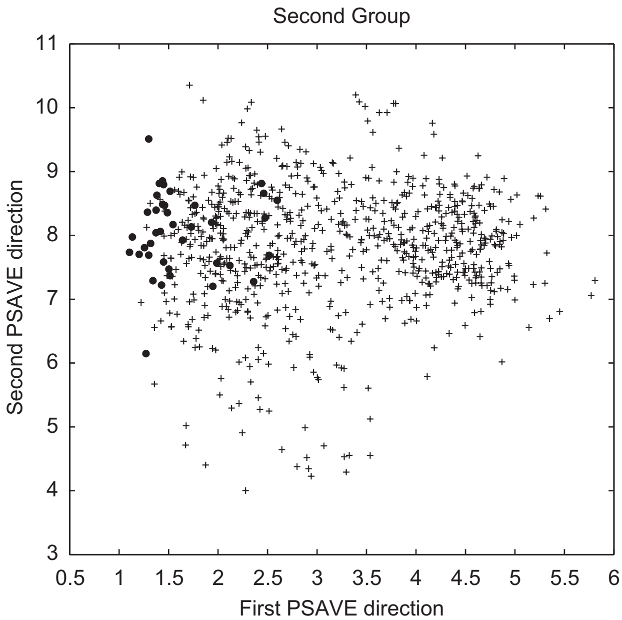

Fig. 3. Plot of the first two directions of the Eurasian watermilfoil data, where "•" denotes that the lake has been infested by water milfoil, "+" denotes that the lake has not been infested.

Table 4 contains the results from applications of the marginal coordinate test to each continuous predictor. In the case of distance, the null hypothesis is that Milfoil is independent of distance given all the other predictors. A 5% backward elimination procedure suggested that Acres, Maxd, PctLitt, SDF, Secchi can be removed without significant loss of information. This was supported by the test that Milfoil is independent of Acres, Maxd, PctLitt, SDF, Secchi given the remaining two continuous predictors distance, TALK, and group variable LakeClass; the p-value for this hypothesis is 0.19. Figs. 4 and 5 show plots of the two significant predictors distance and TALK, marked by the response Milfoil. The plots are very similar to the plots of the two significant directions in Figs. 2 and 3. The infested lakes tend to be closer to other infested lakes, and they tend to have a smaller

Table 4
Marginal coordinate test results for the Eurasian watermilfoil data based on partial SAVE.

Predictor	$T_n(\widehat{H})$	df	p-Value
Distance	43.92	2	0.000
Acres	3.99	2	0.136
Maxd	3.59	2	0.166
PctLitt	0.72	2	0.698
SDF	0.15	2	0.930
Secchi	10.51	2	0.005
TALK	25.69	2	0.000

df: degree of freedoms.

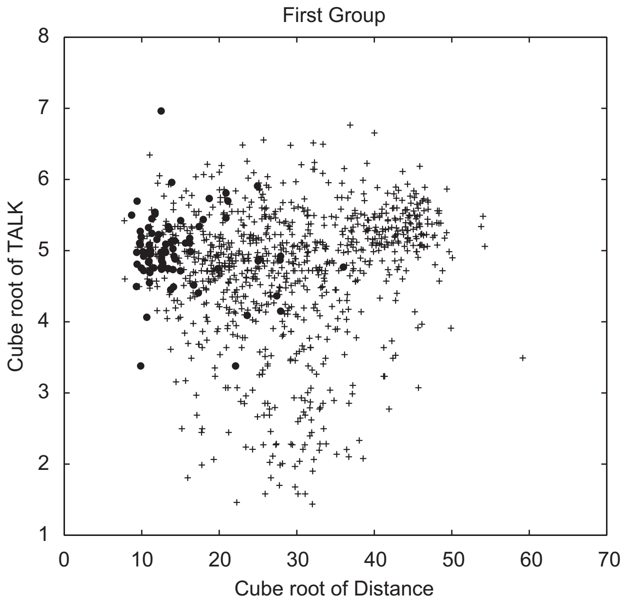

Fig. 4. Plot of the two significant predictors of the Eurasian watermilfoil data, where "•" denotes that the lake has been infested by water milfoil, "+" denotes that the lake has not been infested.

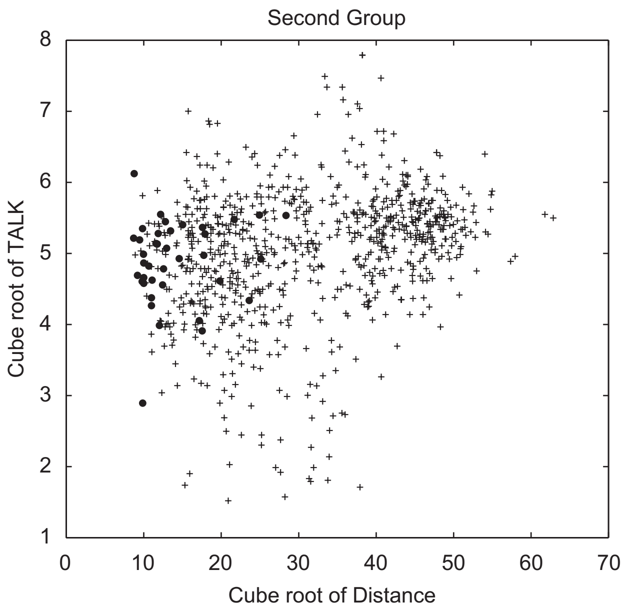

Fig. 5. Plot of the two significant predictors of the Eurasian watermilfoil data, where "•" denotes that the lake has been infested by water milfoil, "+" denotes that the lake has not been infested.

variability in alkalinity. The mean of the cube root of distance for the uninfested lakes is 29.9, while the mean for the infested lakes is only 14.4. This may be because the milfoil is generally spread by boaters, so lakes in close proximity to other infested lakes are probably more at risk. The variance of the cube root of TALK for the uninfested lakes is 0.895, while the variance for the infested lakes is only 0.273. This may be because the conditions are not favorable for milfoil when the alkalinity is too high or too low.

7. Discussion

In this paper we extend the SAVE methodology to regressions with categorical predictors. The proposed method is more comprehensive than methods based on inverse mean, like partial SIR, partial IRE, and general partial SIR. We also constructed marginal coordinate test based on our method. As illustrated in the Eurasian watermilfoil data example, by combining marginal dimension tests and marginal coordinate tests, we can achieve dimension reduction both by removing predictors and by replacing the remaining predictors with linear combination of them, without specifying a parsimonious model.

Softwareforthemethodsdescribedinthisarticleareimplementedinthe“dr” package for R available from cran.r-project.org.

Acknowledgments

The authors greatly appreciate the referees and the editors for their useful comments. R.D. Cook and Y. Shao were supported by Grants DMS-0405360 and DMS-0704098 from the National Science Foundation.

Appendix A. Technical proofs

Lemma 1. *Suppose H is an orthonormal basis of the complement of $\Sigma_{\text{pool}}^{-1/2}\mathscr{V}$, then under the null hypothesis $\mathscr{S}_{Y|X}^{(W)} \subseteq \mathscr{V}$, $T_n(H)$ converges in distribution to $\chi^2[(h-K)(p-m)(p-m+1)]$.*

Proof. First we have

$$T_n(H) = \sum_{w=1}^{K} T_{n,w}(H),$$

where

$$T_{n,w}(H) = \frac{1}{2}\sum_{k=1}^{h_w} n_{wk}\,\text{tr}[(H^{\text{T}}(I_p - \widehat{\text{cov}}(Z|W=w, Y=k))H)^2].$$

By Theorem 1 of Shao et al. (2007) we know that $T_{n,w}(H)$ converges to $\chi^2_{(h_w-1)(p-m)(p-m+1)}$ in distribution as $n_w \to \infty$, which is the same as $n \to \infty$. Since all the observations are iid, $T_{n,w}(H)$ are independent for different w. Therefore $T_n(H)$ converges in distribution to a χ^2 distribution with

$$df = \sum_{w=1}^{K}(h_w - 1)(p-m)(p-m+1) = (h-K)(p-m)(p-m+1). \qquad \square$$

Proof of Theorem 2. First we have

$$T_n(\widehat{H}) = \sum_{w=1}^{K} T_{n,w}(\widehat{H}).$$

Since $\widehat{\Sigma}_{\text{pool}} = \Sigma_{\text{pool}} + \text{O}_{\text{p}}(n^{-1/2})$. Therefore $\widehat{H}\widehat{H}^{\text{T}} = HH^{\text{T}} + \text{O}_{\text{p}}(n^{-1/2})$. By Lemma A1 of Shao et al. (2007),

$$T_{n,w}(\widehat{H}) = T_{n,w}(H) + \text{o}_{\text{p}}(1).$$

Therefore

$$T_n(\widehat{H}) = T_n(H) + \text{o}_{\text{p}}(1).$$

Hence by Lemma 1 $T_n(\widehat{H})$ converges in distribution to $\chi^2[(h-K)(p-m)(p-m+1)]$. $\square$

Proof of Theorem 1. Since $\widehat{M}^{(W)} = M^{(W)} + \mathrm{O_p}(n^{-1/2})$, by Tyler (1981, Lemma 2.1), $\widehat{\theta}\widehat{\theta}^{\mathrm{T}} = \theta\theta^{\mathrm{T}} + \mathrm{O_p}(n^{-1/2})$. Again, by Lemma A1 of Shao et al. (2007),

$$T_{n,w}(\widehat{\theta}) = T_{n,w}(\theta) + \mathrm{o_p}(1).$$

Therefore

$$T_n(\widehat{\theta}) = T_n(\theta) + \mathrm{o_p}(1).$$

Hence by Lemma 1 $T_n(\widehat{H})$ converges in distribution to $\chi^2[(h-K)(p-m)(p-m+1)]$. □

References

Chiaromonte, F., Cook, R.D., Li, B., 2002. Sufficient dimension reduction in regressions with categorical predictors. Ann. Statist. 30, 475–497.
Cook, R.D., 1994. On the interpretation of regression plots. J. Amer. Statist. Assoc. 89, 177–189.
Cook, R.D., 2004. Testing predictor contributions in sufficient dimension reduction. Ann. Statist. 32, 1062–1092.
Cook, R.D., Lee, H., 1999. Dimension reduction in regressions with a binary response. J. Amer. Statist. Assoc. 94, 1187–1200.
Cook, R.D., Ni, L., 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Amer. Statist. Assoc. 100, 410–428.
Cook, R.D., Weisberg, S., 1991. Discussion on `sliced inverse regression for dimension reduction'. J. Amer. Statist. Assoc. 86, 328–332.
Cook, R.D., Weisberg, S., 1999. Applied Regression Including Computing and Graphics. Wiley, New York.
Li, B., Wang, S., 2007. On directional regression for dimension reduction. J. Amer. Statist. Assoc. 102, 997–1008.
Li, K.-C., 1991. Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86, 316–342.
Ni, L., Cook, R.D., 2006. Sufficient dimension reduction in regressions across heterogeneous subpopulations. J. Roy. Statist. Soc. Ser. B 68, 89–107.
Schupp, D., 1992. An ecological classification of Minnesota lakes with associated fish communities. Minnesota Department of Natural Resources, Investigational Report 417.
Shao, Y., Cook, R.D., Weisberg, S., 2007. Marginal tests with sliced average variance estimation. Biometrika 94, 285–296.
Smith, C.S., Barko, J.W., 1990. Ecology of eurasian watermilfoil. J. Aquatic Plant Management 28, 55–64.
Tyler, D., 1981. Asymptotic inference for eigenvectors. Ann. Statist. 9, 725–736.
Wen, X., Cook, R.D., 2007. Optimal sufficient dimension reduction in regressions with categorical predictors. J. Statist. Plann. Inference 137, 1961–1978.