The Korean Journal of Applied Statistics (2010)
23*(5), 977–985*
DOI: 10.5351/KJAS.2010.23.5.977

Integrated Partial Sufficient Dimension Reduction with Heavily Unbalanced Categorical Predictors

Jae Keun Yoo[1]

[1]Department of Statistics, Ewha Womans University

(Received July 2010; accepted September 2010)

Abstract

In this paper, we propose an approach to conduct partial sufficient dimension reduction with heavily unbalanced categorical predictors. For this, we consider integrated categorical predictors and investigate certain conditions that the integrated categorical predictor is fully informative to partial sufficient dimension reduction. For illustration, the proposed approach is implemented on optimal partial sliced inverse regression in simulation and data analysis.

Keywords: Integration, partial dimension subspaces, sufficient dimension reduction, regression, unbalanced categorical predictors.

1. Introduction

Sufficient dimension reduction(SDR) in regression of $Y|\mathbf{X} \in \mathbb{R}^p$ replaces the original p-dimensional many-valued or continuous predictors $\mathbf{X}$ by a lower-dimensional linear projection predictor without loss of information about selected aspects of the conditional distribution of $Y|\mathbf{X}$. Equivalently, SDR pursues to find the minimal subspaces $\mathcal{S}_{f(\mathbf{X})}$ satisfying

$$Y \perp\!\!\!\perp f(\mathbf{X})|\mathbf{P}_{\mathcal{S}_{f(\mathbf{X})}}\mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ stands for independence, $f(\mathbf{X})$ characterizes the selected aspects of $Y|\mathbf{X}$, $\mathbf{P}_{\mathcal{S}}$ represents the orthogonal projection onto a subspace $\mathcal{S}$ with usual inner-product space, and $\dim\{\mathcal{S}_{f(\mathbf{X})}\} \leq p$. We call the lower-dimensional linear projection predictor $\mathbf{P}_{\mathcal{S}_{f(\mathbf{X})}}\mathbf{X}$ *sufficient predictors*.

The form of $f(\mathbf{X})$ in (1.1) is determined according to the selected aspects of $Y|\mathbf{X}$. If the conditional distribution itself is of main interest in regression, $f(\mathbf{X})$ becomes $\mathbf{X}$. And, such minimal subspace is called the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ (Cook, 1998). Then $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ can replace $\mathbf{X}$ without loss of information on $Y|\mathbf{X}$. If the main interest is placed on the first conditional moment $E(Y|\mathbf{X})$, then $f(\mathbf{X})$ becomes $E(Y|\mathbf{X})$, and the related minimal subspace is called the *central mean subspace* $\mathcal{S}_{E(Y|\mathbf{X})}$ (Cook and Li, 2002). In this case, $\mathbf{P}_{\mathcal{S}_{E(Y|\mathbf{X})}}\mathbf{X}$ can replace $\mathbf{X}$ without loss of information on

For the corresponding author Jae Keun Yoo, this work was supported by Basic Science Research Program through the National Research Foundation of Korea(KRF) funded by the Ministry of Education, Science and Technology(2010-0003189).

[1]Assistant Professor, Department of Statistics, Ewha Womans University, Seoul 120-750, Republic of Korea. E-mail: peter.yoo@ewha.ac.kr

$E(Y|\mathbf{X})$. When the primary focus is given in the first k conditional moments of $Y|\mathbf{X}$, $f(\mathbf{X})$ is equal to $\{E(Y|\mathbf{X}), M^{(2)}(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})\}$, and we call the minimal space the *central* k^{th}*-moment subspace* $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ (Yin and Cook, 2002), where $M^{(k)}(Y|\mathbf{X}) = E[\{Y - E(Y|\mathbf{X})\}^k|\mathbf{X}]$ for $k \geq 2$. In this context, $\mathbf{P}_{\mathcal{S}^{(k)}_{Y|\mathbf{X}}}\mathbf{X}$ can replace $\mathbf{X}$ without loss of information on the first k conditional moments.

When a categorical predictor W with c levels is involved in regression, the definition of SDR introduced above is not directly applicable. In such case, we pursue partial SDR to reduce dimensions of many-valued or continuous predictors $\mathbf{X}$ alone conditioning W. Then a related conditional independence statement for partial SDR is as follows:

$$Y \perp\!\!\!\perp f(\mathbf{X}, W)|(\mathbf{P}_{\mathcal{S}_{f(\mathbf{X},W)}}\mathbf{X}, W), \tag{1.2}$$

Similarly to (1.1), in partial SDR context, the related minimal subspaces satisfying (1.2) are called the *partial central subspace* $\mathcal{S}^{W}_{Y|\mathbf{X}}$ (Chiaromonte *et al.*, 2002), the *partial central mean subspace* $\mathcal{S}^{W}_{E(Y|\mathbf{X})}$ (Li *et al.*, 2003) the *partial* k^{th}*-moment subspace* $\mathcal{S}^{W(k)}_{Y|\mathbf{X}}$ (Yoo, 2009) depending on the selected aspects of $Y|\mathbf{X}$, and the corresponding forms of $f(\mathbf{X}, W)$ are $\mathbf{X}$, $E(Y|\mathbf{X}, W)$ and $\{E(Y|W, \mathbf{X}), M^{(2)}(Y|\mathbf{X}, W), \ldots, M^{(k)}(Y|\mathbf{X}, W)\}$ respectively, where $M^{(k)}(Y|\mathbf{X}, W) = E[\{Y - E(Y|\mathbf{X}, W)\}^k|\mathbf{X}, W]$ for $k \geq 2$.

When two or more categorical predictors $\mathbf{W} = (W_1, \ldots, W_q)$ with $q \geq 2$ are involved, we need to do a full hierarchical categorization of $\mathbf{W}$. For example, supposing that two categorical predictors of W_1 and W_2 are involved in regression and each has two levels of 0 and 1, we can replace them by a new categorical variable W_{FH} with four levels: $W_{\text{FH}} = 0$, if $(W_1, W_2) = (0, 0)$; $W_{\text{FH}} = 1$, if (0,1); $W_{\text{FH}} = 2$, if (1,0); $W_{\text{FH}} = 3$, if (1,1). We will call such W_{FH} as a fully hierarchically categorized predictor. Considering q categorical predictors with each c_i levels, $i = 1, \ldots, q$, W_{FH} will have totally $\Pi_{i=1}^{q} c_i$ levels.

One issue possibly arising with many categorical predictors is heavily unbalanced within some levels of W_{FH}. In practice, this must be cause of concern in methodological application for partial SDR due to too few data in some categories of W_{FH}.

This article suggest an approach to conduct partial SDR under heavily unbalanced categorical predictors. For this, we will take integration of some levels of W_{FH}, and define integrated hierarchical predictor W_{IH}. We will investigate conditions that W_{IH} is fully informative to partial SDR. Then, with W_{IH} under the conditions, usual partial SDR methodologies can be directly applied for dimension reduction of many-valued or continuous predictors.

To avoid interrupting the discussion, several notations are defined. A set of many-valued and continuous predictors will be denoted as $\mathbf{X} = (X_1, \ldots, X_p)^{\text{T}}$, while a set of categorical predictors will be defined as $\mathbf{W} = (W_1, \ldots, W_q)^{\text{T}}$. We assume throughout that the data $(Y_i, \mathbf{X}_i, \mathbf{W}_i)$ are a random sample for $(Y, \mathbf{X}, \mathbf{W})$ for $i = 1, 2, \ldots, n$. A generic pair of $(Y_w, \mathbf{X}_w)$ indicates a subpopulation of $(Y, \mathbf{X})$ such that $(Y, \mathbf{X})|W = w$, and a regression of $Y_w|\mathbf{X}_w$, equivalently, $Y|(\mathbf{X}, W = w)$, will be called a conditional regression within the subpopulation of $W = w$. A notation of $\mathcal{S}(\mathbf{B})$ stands for as a subspace of $\mathbb{R}^p$ spanned by the columns of a $p \times r$ matrix $\mathbf{B}$.

2. Integrated Categorization

2.1. Integration of categorical predictors

Consider a regression of $Y|(\mathbf{X} \in \mathbb{R}^p, \mathbf{W})$, where $\mathbf{W} = (W_1, \ldots, W_q)^{\text{T}}$ and $q \geq 2$. Now, we assume that, in one of W_is, at least, data is heavily unbalanced. That is, certain levels of W_is contain few

observations. Therefore, some levels of W_{FH} will have even fewer or no observations. For illustration purposes, we consider $\mathbf{W}$ and W_{FH} demonstrated in Section 1: $\mathbf{W} = (W_1, W_2)$ with each having two levels; $W_{\mathrm{FH}} = 0$, if $(W_1, W_2) = (0, 0)$; $W_{\mathrm{FH}} = 1$, if (0,1); $W_{\mathrm{FH}} = 2$, if (1,0); $W_{\mathrm{FH}} = 3$, if (1,1).

Here suppose that the number of observations for $W_1 = 1$ is too small. Under this circumstances usual application of partial SDR methods introduced in Section 1 is not plausible.

To overcome this deficit, as an alternative, we meaningfully integrate levels of W_{FH} to have few observations and construct an integrated categorical predictor W_{IH}. In the example above, the following W_{IH} can be considered: $W_{\mathrm{IH}} = 0$, if $(W_1, W_2) = (0, 0)$; $W_{\mathrm{IH}} = 1$, if (0,1); $W_{\mathrm{IH}} = 2$, if $W_1 = 1$. In W_{IH}, we can easily see that the two levels of $W_{\mathrm{FH}} = 2$ and $W_{\mathrm{FH}} = 3$ in W_{FH} are combined into one level.

To facilitate understanding of W_{FH} and W_{IH}, we partition W_{FH} into two parts of $W_{\mathrm{FH}}^{(N)}$ and $W_{\mathrm{FH}}^{(I)}$. The former $W_{\mathrm{FH}}^{(N)}$ is parts of W_{FH} not integrated to make W_{IH}, while the latter $W_{\mathrm{FH}}^{(I)}$ corresponds to parts of W_{FH} integrated for W_{IH}. In the example, $W_{\mathrm{FH}}^{(N)}$ contains $W_{\mathrm{FH}} = 0$ and $W_{\mathrm{FH}} = 1$, and $W_{\mathrm{FH}}^{(I)}$ does $W_{\mathrm{FH}} = 2$ and $W_{\mathrm{FH}} = 3$.

The construction of W_{IH} is done by integrating $W_{\mathrm{FH}}^{(I)}$ with keeping $W_{\mathrm{FH}}^{(N)}$. Let a categorical predictor formed by integrating $W_{\mathrm{FH}}^{(I)}$ be $W_{\mathrm{IH}}^{(I)}$. In the example, $W_{\mathrm{IH}}^{(I)}$ corresponds to $W_{\mathrm{IH}} = 2$. By this, W_{IH} can be partitioned into two parts of $W_{\mathrm{FH}}^{(N)}$ and $W_{\mathrm{IH}}^{(I)}$.

Using these notations, we can re-write W_{FH} and W_{IH} as follows: $W_{\mathrm{FH}} = \{W_{\mathrm{FH}}^{(N)}, W_{\mathrm{FH}}^{(I)}\}^{\mathrm{T}}$ and $W_{\mathrm{IH}} = \{W_{\mathrm{FH}}^{(N)}, W_{\mathrm{IH}}^{(I)}\}^{\mathrm{T}}$.

It should be noted that $W_{\mathrm{IH}}^{(I)}$ is a special case of $W_{\mathrm{FH}}^{(I)}$, in the sense that $W_{\mathrm{IH}}^{(I)}$ is constructed by integrating $W_{\mathrm{FH}}^{(I)}$. Therefore, theoretically, usage of $W_{\mathrm{FH}}^{(I)}$ is more informative than that of $W_{\mathrm{IH}}^{(I)}$ in partial SDR. In the next section, we investigate conditions to guarantee that W_{IH} is as informative to partial SDR as W_{FH}.

2.2. Required conditions

First we define target subspaces $\mathcal{S}^{\mathrm{FH}}_{f(\mathbf{X}_w)}$ for $f(\mathbf{X}_w)$ for the conditional regression of $Y_w|\mathbf{X}_w$, which is a regression $Y|\mathbf{X}$ within a subpopulation $W_{\mathrm{FH}} = w$. Similarly, we define that $\mathcal{S}^{\mathrm{FH}(N)}_{f(\mathbf{X}_{w_{\mathrm{n}}})}$, $\mathcal{S}^{\mathrm{FH}(I)}_{f(\mathbf{X}_{w_{\mathrm{i}}})}$ and $\mathcal{S}^{\mathrm{IH}(I)}_{f(\mathbf{X}_{w'_{\mathrm{i}}})}$ are target subspaces for $W_{\mathrm{FH}}^{(N)} = w_{\mathrm{n}}$, $W_{\mathrm{FH}}^{(I)} = w_{\mathrm{i}}$ and $W_{\mathrm{IH}}^{(I)} = w'_{\mathrm{i}}$ respectively. Then the following relations trivially hold among $\mathcal{S}_{f(\mathbf{X}_w)}$, $\mathcal{S}^{\mathrm{FH}(N)}_{f(\mathbf{X}_{w_{\mathrm{n}}})}$, $\mathcal{S}^{\mathrm{FH}(I)}_{f(\mathbf{X}_{w_{\mathrm{i}}})}$ and $\mathcal{S}^{\mathrm{IH}(I)}_{f(\mathbf{X}_{w'_{\mathrm{i}}})}$:

$$\oplus_{w_{\mathrm{i}'}=1}^{c'_{\mathrm{i}}} \mathcal{S}^{\mathrm{IH}(I)}_{f(\mathbf{X}_{w'_{\mathrm{i}}})} \subseteq \oplus_{w_{\mathrm{i}}=1}^{c^*_{\mathrm{i}}} \mathcal{S}^{\mathrm{FH}(I)}_{f(\mathbf{X}_{w_{\mathrm{i}}})}, \tag{2.1}$$

$$\begin{aligned} \mathcal{S}^{W_{\mathrm{IH}}}_{f(\mathbf{X})} &:= \left\{\oplus_{w_{\mathrm{n}}=1}^{c^*_{\mathrm{n}}} \mathcal{S}^{\mathrm{FH}(N)}_{f(\mathbf{X}_{w_{\mathrm{n}}})}\right\} \oplus \left\{\oplus_{w'_{\mathrm{i}}=1}^{c'_{\mathrm{i}}} \mathcal{S}^{\mathrm{IH}(I)}_{f(\mathbf{X}_{w'_{\mathrm{i}}})}\right\} \\ &\subseteq \left\{\oplus_{w_{\mathrm{n}}=1}^{c^*_{\mathrm{n}}} \mathcal{S}^{\mathrm{FH}(N)}_{f(\mathbf{X}_{w_{\mathrm{n}}})}\right\} \oplus \left\{\oplus_{w_{\mathrm{i}}=1}^{c^*_{\mathrm{i}}} \mathcal{S}^{\mathrm{FH}(I)}_{f(\mathbf{X}_{w_{\mathrm{i}}})}\right\} = \oplus_{w=1}^{c^*} \mathcal{S}^{\mathrm{FH}}_{f(\mathbf{X}_w)} =: \mathcal{S}^{W_{\mathrm{FH}}}_{f(\mathbf{X})}, \end{aligned} \tag{2.2}$$

where the notation $\oplus$ indicates the direct sum among subspaces ($\mathcal{S}_1 \oplus \mathcal{S}_2 = v_1 + v_2; v_1 \in \mathcal{S}_1, v_2 \in \mathcal{S}_2$) and c^*, c^*_{n}, c^*_{i} and c'_{i} represent total levels of W_{FH}, $W_{\mathrm{FH}}^{(N)}$, $W_{\mathrm{FH}}^{(I)}$ and $W_{\mathrm{IH}}^{(I)}$ respectively.

Statement (2.2) directly comes from (2.1). Then, under partial SDR contexts, the following relation is established for ultimate target partial dimension reduction subspaces $\mathcal{S}^{\mathbf{W}}_{f(\mathbf{X})}$ for $f(\mathbf{X})$ in the regression of $Y|(\mathbf{X}, \mathbf{W})$:

$$\mathcal{S}^{W_{\mathrm{IH}}}_{f(\mathbf{X})} \subseteq \mathcal{S}^{W_{\mathrm{FH}}}_{f(\mathbf{X})} = \mathcal{S}^{\mathbf{W}}_{f(\mathbf{X})}. \tag{2.3}$$

Through the relations of $W_{\text{FH}}^{(N)}$, $W_{\text{FH}}^{(I)}$ and $W_{\text{IH}}^{(I)}$, conditions to guarantee that $\mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}} = \mathcal{S}_{f(\mathbf{X})}^{W_{\text{IH}}}$ are summarized in the next lemma:

Lemma 2.1 *The equivalence of* $\mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}} = \mathcal{S}_{f(\mathbf{X})}^{W_{\text{IH}}}$ *is established under either of the following conditions:*

$$1.\ \oplus_{w_{\text{i}}=1}^{c_{\text{i}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}})}^{\text{FH}(I)} \subseteq \oplus_{w_{\text{n}}=1}^{c_{\text{n}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{n}}})}^{\text{FH}(N)}; \qquad 2.\ \oplus_{w_{\text{i}}=1}^{c_{\text{i}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}})}^{\text{FH}(I)} = \oplus_{w_{\text{i}}'=1}^{c_{\text{i}}'} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}'})}^{\text{IH}(I)}.$$

Proof. (Condition 1) Under condition 1, $\mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}} = \oplus_{w_{\text{n}}=1}^{c_{\text{n}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{n}}})}^{\text{FH}(N)} \subseteq \mathcal{S}_{f(\mathbf{X})}^{W_{\text{IH}}}$, and hence we have $\mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}} = \mathcal{S}_{f(\mathbf{X})}^{W_{\text{IH}}}$. (Condition 2) By condition 2, we have $\mathcal{S}_{f(\mathbf{X})}^{W_{\text{IH}}} = \{\oplus_{w_{\text{n}}=1}^{c_{\text{n}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{n}}})}^{\text{FH}(N)}\} \oplus \{\oplus_{w_{\text{i}}'=1}^{c_{\text{i}}'} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}'})}^{\text{IH}(I)}\}$ $= \{\oplus_{w_{\text{n}}=1}^{c_{\text{n}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{n}}})}^{\text{FH}(N)}\} \oplus \{\oplus_{w_{\text{i}}=1}^{c_{\text{i}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}})}^{\text{FH}(I)}\} = \mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}}$. This completes the proof. □

The conditions 1, 2 in Lemma 2.1 do not necessarily imply each other. To see that one of the conditions is at least satisfied, one simple way is to conduct partial SDR for each categorical predictor W_i, $i = 1, \ldots, q$ and to study how strongly the estimated sufficient predictors are correlated. The partial SDR for the full regression of $Y|(\mathbf{X}, \mathbf{W})$ pursues to find $\boldsymbol{\eta}$ such that $Y \perp\!\!\!\perp f(\mathbf{X}, \mathbf{W})|(\mathbf{P}_{\mathcal{S}(\boldsymbol{\eta})}\mathbf{X}, \mathbf{W})$. This statement directly implies that $Y \perp\!\!\!\perp f(\mathbf{X}, W_i)|(\mathbf{P}_{\mathcal{S}(\boldsymbol{\eta})}\mathbf{X}, W_i)$, $i = 1, \ldots, q$. High correlations among the estimated sufficient predictors acquired from $Y|(\mathbf{X}, W_i)$, $i = 1, \ldots, q$, implies that common sufficient predictors, saying that $\mathbf{P}_{\mathcal{S}(\hat{\boldsymbol{\eta}})}\mathbf{X}$ should be enough for all regressions of $Y|(\mathbf{X}, W_i)$. Then we may, in practice, expect that the statement of $Y \perp\!\!\!\perp f(\mathbf{X}, \mathbf{W})|(\mathbf{P}_{\mathcal{S}(\boldsymbol{\eta})}\ \mathbf{X}, \mathbf{W})$ will hold for the common sufficient predictors $\mathbf{P}_{\mathcal{S}(\hat{\boldsymbol{\eta}})}\mathbf{X}$. Then we can attain the full partial SDR with some of W_is, which form $W_{\text{FH}}^{(N)}$. This implies that $\oplus_{w_{\text{n}}=1}^{c_{\text{n}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{n}}})}^{\text{FH}(N)} = \mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}}$, and hence condition 1 seems satisfied. We apply this argument to a subset of of $\mathbf{W}$ used to construct $W_{\text{IH}}^{(I)}$, saying that $\mathbf{W}' = (W_1', \ldots, W_r')$ with $r \leq q$. If the estimated sufficient predictors obtained from partial SDRs of $Y|(\mathbf{X}, W_i')$s are highly correlated, it is expected that $\mathcal{S}_{f(\mathbf{X})}^{\mathbf{W}'} = \oplus_{w_{\text{i}}=1}^{c_{\text{i}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}})}^{\text{FH}(I)} = \oplus_{w_{\text{i}}'=1}^{c_{\text{i}}'} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}'})}^{\text{IH}(I)}$, and hence condition 2 seems satisfied.

Since, under condition 1, all information for the target subspaces are given in the conditional regressions of $Y_{w_{\text{n}}}|\mathbf{X}_{w_{\text{n}}}$ within subpopulations of $W_{\text{FH}}^{(N)}$, $w_{\text{n}} = 1, \ldots, c_{\text{n}}^*$, and they have a major portion of observations, it is practically expected to produce more accurate dimension reduction under condition 1 than under condition 2. In this case, few observations in $W_{\text{FH}}^{(I)}$ will not matter, because, regardless of $W_{\text{IH}}^{(I)}$, $\mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}}$ will be well-estimated through $\oplus_{w_{\text{n}}=1}^{c_{\text{n}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{n}}})}^{\text{FH}(N)}$.

For condition 2, the information of the conditional regressions of $Y_{w_{\text{i}}}|\mathbf{X}_{w_{\text{i}}}$ within subpopulations of $W_{\text{FH}}^{(I)}$ must be essential for good estimation of $\mathcal{S}_{f(\mathbf{X})}^{W_{\text{FH}}}$. Condition 2, in theory, guarantee full coverage of $\mathcal{S}_{Y|\mathbf{X}}^{W_{\text{FH}}}$ through $\oplus_{w_{\text{i}}=1}^{c_{\text{i}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}})}^{\text{FH}(I)}$ estimated by $W_{\text{IH}}^{(I)}$, but, the estimation by $W_{\text{IH}}^{(I)}$ may still struggle due to lack of observations in $W_{\text{FH}}^{(I)}$ in practice. Therefore, it is important to construct W_{IH} so that $W_{\text{IH}}^{(I)}$ has good sample sizes to well-estimate $\oplus_{w_{\text{i}}=1}^{c_{\text{i}}^*} \mathcal{S}_{f(\mathbf{X}_{w_{\text{i}}})}^{\text{FH}(I)}$.

3. Simulation and Case Study

For illustration purposes, we adopt optimal partial sliced inverse regression(OPSIR; Wen and Cook, 2007), which can be applied in heterogeneous predictor covariances across the subpopulations by categorical predictors and estimate the partial central subspace $\mathcal{S}_{f(\mathbf{X})}^{\mathbf{W}} := \mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}}$. A notation d will represent the structural dimension of $\mathcal{S}_{Y|\mathbf{X}}^{\mathbf{W}}$ throughout the rest of the paper.

Table 3.1. Estimation of $\mathcal{S}^{\mathbf{W}}_{Y|\mathbf{X}}$ for Model 1 in Section 3.1

	W_{FH}		W_{IH}	
	Percents of $\hat{d}=1$	$\bar{r}$	Percents of $\hat{d}=1$	$\bar{r}$
$n=100$	60.3	0.969	84.5	0.982
$n=200$	69.4	0.992	87.8	0.993
$n=300$	78.8	0.998	90.4	0.998

In the simulation, we generated five dimensional predictors $\mathbf{X}=(X_1,\ldots,X_5)^{\mathrm{T}}$ from either $N(0,\mathbf{I}_5)$ or $N(0,\mathbf{\Sigma})$, where $\mathbf{I}_p\in\mathbb{R}^{p\times p}$ represents the identity matrix, and $\mathbf{\Sigma}_{i,j}=1$ for $i=j$ and $\mathbf{\Sigma}_{i,j}=0.25$ for $i\neq j$, $i=1,\ldots,5$ and $j=1,\ldots,5$. For dimension determination, nominal level 5% was used for all simulations.

3.1. Model 1

Model 1 was constructed under condition 1 to compare estimation performance of $\mathcal{S}^{\mathbf{W}}_{Y|\mathbf{X}}$ using W_{FH} and W_{IH}. We considered two categorical predictors of W_1 and W_2 with each having two levels. Therefore, we defined W_{FH} as follows: $W_{\mathrm{FH}}=0$, if $(W_1,W_2)=(0,0)$; $W_{\mathrm{FH}}=1$, if $(W_1,W_2)=(0,1)$; $W_{\mathrm{FH}}=2$, if $(W_1,W_2)=(1,0)$; $W_{\mathrm{FH}}=3$, if $(W_1,W_2)=(1,1)$. Let n be the total sample size and n_{ij} be the sample sizes with $(W_1,W_2)=(i,j)$, $i=0,1$ and $j=0,1$. Then we set $n_{00}=0.4*n$, $n_{01}=0.4*n$, $n_{10}=0.1*n$ and $n_{11}=0.1*n$.

We randomly sampled predictors $\mathbf{X}$ within each category of W_{FH} as follows: if $W_{\mathrm{FH}}=0$, $\mathbf{X}_{00}\sim N(0,\mathbf{I}_5)$; if $W_{\mathrm{FH}}=1$, $\mathbf{X}_{01}\sim N(0,\mathbf{\Sigma})$; if $W_{\mathrm{FH}}=2$, $\mathbf{X}_{10}\sim N(0,\mathbf{I}_5)$; if $W_{\mathrm{FH}}=3$, $\mathbf{X}_{11}\sim N(0,\mathbf{\Sigma})$.

Next the following regressions were considered: if $W_{\mathrm{FH}}=0$, $Y_{00}=X_{00_1}+X_{00_2}+0.1*\varepsilon_{00}$; if $W_{\mathrm{FH}}=1$, $Y_{01}=\exp\{0.5*(X_{01_1}+X_{01_2})\}+0.1*\varepsilon_{01}$; if $W_{\mathrm{FH}}=2$, $Y_{10}=X_{10_1}+X_{10_2}+0.1*\varepsilon_{10}$; if $W_{\mathrm{FH}}=3$, $Y_{11}=\exp\{0.5*(X_{11_1}+X_{11_2})\}+0.1*\varepsilon_{11}$; where $\varepsilon_{i^*j^*}\overset{iid}{\sim}N(0,1)\perp\!\!\!\perp\mathbf{X}_{ij}$ for $i=0,1$, $j=0,1$, $i^*=0,1$ and $j^*=0,1$.

By construction, within all levels of W_{FH}, all conditional regressions depend on $\mathbf{X}$ only through X_1+X_2, and hence condition 1 holds; and, $\mathcal{S}^{\mathbf{W}}_{Y|\mathbf{X}}$ is spanned by $\boldsymbol{\eta}=(1,1,0,0,0)^{\mathrm{T}}$. Since there were fewer observations in $W_{\mathrm{FH}}=2$ and $W_{\mathrm{FH}}=3$, we constructed W_{IH} by integrating $W_{\mathrm{FH}}=2$ and 3: $W_{\mathrm{IH}}=0$, if $(W_1,W_2)=(0,0)$; $W_{\mathrm{IH}}=1$, if $(W_1,W_2)=(0,1)$; $W_{\mathrm{IH}}=2$, otherwise.

In the simulation, we used two slices for each conditional regression of $Y_w|\mathbf{X}_w$ within $W_{\mathrm{FH}}=w$, $w=0,1,2,3$ and for $Y_{w'}|\mathbf{X}_{w'}$ within $W_{\mathrm{IH}}=w'$, $w'=0,1,2$. As summaries of the simulation study, we report percentages of $\hat{d}=1$ for the structural dimension estimation and the averages $\bar{r}_1$ of $|r|=|\sqrt{R^2}|$ computed from the OLS fits of X_1+X_2 on $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ to measure how well the true basis $\boldsymbol{\eta}$ is estimated. Table 3.1 summaries the results for n =100, 200 and 300. The percentages for the decisions that $\hat{d}=0$ were zeros (not reported).

According to Table 3.1, W_{IH} resulted in better estimation of $\mathcal{S}^{\mathbf{W}}_{Y|\mathbf{X}}$ than W_{FH} for all sample sizes considered. Lack of observations with in $W_{\mathrm{FH}}=2$ and $W_{\mathrm{FH}}=3$ seemed to produce poor results of W_{FH} with smaller samples. In either case, the true basis was well-estimated. This simulation confirms the possible advantage of W_{IH} over W_{FH} in practice.

3.2. Model 2

For Model 2, we considered two categorical predictors of W_1 with 2 levels and W_2 with 3 levels. The hierarchically categorized predictor W_{FH} is constructed as follows: $W_{\mathrm{FH}}=0$, if $(W_1,W_2)=(0,0)$;

Table 3.2. Estimation of $\mathcal{S}^{\mathbf{W}}_{Y|\mathbf{X}}$ in the simulation example of Section 3.2

	Percents of $\hat{d}=1$	Percents of $\hat{d}=2$	Percents of $\hat{d}>2$	$\bar{r}_1$	$\bar{r}_2$
$n=100$	38.1	54.3	7.6	0.985	0.710
$n=200$	15.3	77.9	6.8	0.997	0.872
$n=300$	4.1	90.1	5.9	0.998	0.930

$W_{\mathrm{FH}}=1$, if $(W_1,W_2)=(0,1)$; $W_{\mathrm{FH}}=2$, if $(W_1,W_2)=(0,2)$; $W_{\mathrm{FH}}=3$, if $(W_1,W_2)=(1,0)$; $W_{\mathrm{FH}}=4$, if $(W_1,W_2)=(1,1)$; $W_{\mathrm{FH}}=5$, if $(W_1,W_2)=(1,2)$. Let n_{ij} be the sample sizes with $(W_1,W_2)=(i,j)$, $i=0,1$ and $j=0,1,2$. Here we set $n_{00}=0.28*n$, $n_{01}=0.3*n$, $n_{02}=0.3*n$ $n_{10}=0.04*n$, $n_{11}=0.04*n$ and $n_{12}=0.04*n$.

Predictors $\mathbf{X}=(X_1,\ldots,X_5)^{\mathrm{T}}$ were randomly generated within each category of W_{FH}: if $W_{\mathrm{FH}}=0$, $\mathbf{X}_{00}\sim N(0,\mathbf{\Sigma})$; if $W_{\mathrm{FH}}=1$, $\mathbf{X}_{01}\sim N(0,\mathbf{I}_5)$; if $W_{\mathrm{FH}}=2$, $\mathbf{X}_{02}\sim N(0,\mathbf{I}_5)$; if $W_{\mathrm{FH}}=j+3$, $\mathbf{X}_{1j}\sim N(0,\mathbf{\Sigma})$, $j=0,1,2$.

Finally the following regressions were constructed: if $W_{\mathrm{FH}}=0$, $Y_{00}=X_{00_1}+0.1*\varepsilon_{00}$; if $W_{\mathrm{FH}}=1$, $Y_{01}=X_{01_1}+0.1*\varepsilon_{01}$; if $W_{\mathrm{FH}}=2$, $Y_{02}=\exp(X_{02_1})+0.1*\varepsilon_{02}$; if $W_{\mathrm{FH}}=3$, $Y_{10}=X_{10_2}+0.1*\varepsilon_{10}$; if $W_{\mathrm{FH}}=4$, $Y_{11}=X_{11_2}+0.1*\varepsilon_{11}$; if $W_{\mathrm{FH}}=5$, $Y_{12}=\exp(X_{12_2})+0.1*\varepsilon_{12}$, where $\varepsilon_{i^*j^*}\overset{iid}{\sim}N(0,1)\perp\!\!\!\perp\mathbf{X}_{ij}$ for $i=0,1$, $j=0,1,2$, $i^*=0,1$ and $j^*=0,1,2$.

By construction, with $W_{\mathrm{FH}}=0,1,2$, that is marginally $W_1=0$, the regression depends on $\mathbf{X}$ only through X_1, while the regression does on $\mathbf{X}$ only through X_2 with $W_{\mathrm{FH}}=3,4,5$, equivalently marginally $W_1=1$. Therefore, in the example, $\mathcal{S}^{\mathbf{W}}_{Y|\mathbf{X}}$ is spanned by the two columns of $\boldsymbol{\eta}=\{(1,0,0,0,0),(0,1,0,0,0)\}^{\mathrm{T}}$.

It can be easily noted that the simulated data is heavily unbalanced with $W_1=1$, which has only 12% of total samples. Therefore, we construct W_{IH} by integrating $W_{\mathrm{FH}}=3,4$ and 5: $W_{\mathrm{IH}}=0$, if $(W_1,W_2)=(0,0)$; $W_{\mathrm{IH}}=1$, if $(W_1,W_2)=(0,1)$; $W_{\mathrm{IH}}=2$, if $(W_1,W_2)=(0,2)$; $W_{\mathrm{IH}}=3$, otherwise. With this integration, condition 2 in Lemma 2.1 is satisfied, because the conditional regression within subpopulation of $W_{\mathrm{IH}}=3$ depends only through X_2.

In the simulation, we used two slices for each conditional regression of $Y_{w'}|\mathbf{X}_{w'}$ within $W_{\mathrm{IH}}=w'$, $w'=0,1,2,3$. As summaries of the simulation study, we report percentages of $\hat{d}=2$ for dimension estimation and the averages $\bar{r}_1$ and $\bar{r}_2$ of $|r_1|=|\sqrt{R_1^2}|$ and $|r_2|=|\sqrt{R_2^2}|$ computed from the OLS fits of X_1 on $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ and of X_2 on $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ respectively for basis estimation. Table 3.2 summaries the results for n =100, 200 and 300. The percentages for the decisions that $\hat{d}=0$ were zeros (not reported).

For $n=100$, since there are only 12 observations for the conditional regression of $Y_3|\mathbf{X}_3$ within the subpopulation $W_{\mathrm{IH}}=3$, and they have to account for the second direction X_2, it is naturally expected that $\{\mathcal{S}(0,1,0,0,0)^{\mathrm{T}}\}$ will not be accurately estimated and hence the true dimension d will be often underestimated to $\hat{d}=1$. This expectation is confirmed according to Table 3.2. With $n=100$, d is determined to 1 about 40% of the time, while the percentages of the correct decisions are 54%. Relatively low $\bar{r}_2$ (0.710) compared to $\bar{r}_1$ (0.985) also supports this. For $n=200$, however, both the correct decision percentages and $\bar{r}_2$ are greatly improved, at least, by 20% and by 15% respectively, and the dimension and basis estimation are quite reliable with $n=300$. Nominal level 5%, that is percentages of $\hat{d}>2$, is consistently well-estimated regardless of sample sizes.

3.3. Case study: Beta-carotene plasma

For illustration purposes, we investigate a regression study of Beta-carotene plasma concentration

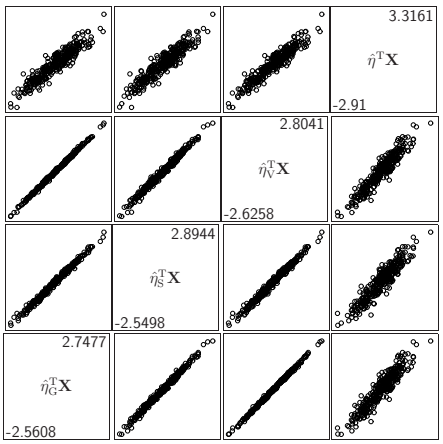


Figure 3.1. A scatter plot matrix of $\hat{\boldsymbol{\eta}}_{\mathrm{G}}^{\mathrm{T}}\mathbf{X}$, $\hat{\boldsymbol{\eta}}_{\mathrm{S}}^{\mathrm{T}}\mathbf{X}$, $\hat{\boldsymbol{\eta}}_{\mathrm{V}}^{\mathrm{T}}\mathbf{X}$ and $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ in Section 3.3

levels given dietary factors, which are number of calories consumed per day (Calorie), grams of fiber consumed per day (Fiber), weight/height^2 (Quetelet), dietary retinol consumed per day (mcg, Ret.diet), and three categorical predictors of Gender (0 = male; 1 = female), Smoke (0 = non smoker; 1 = former smoker; 2 = current smoker) and Vitamin use (0 = often; 1 = sometimes; or 2 = no).

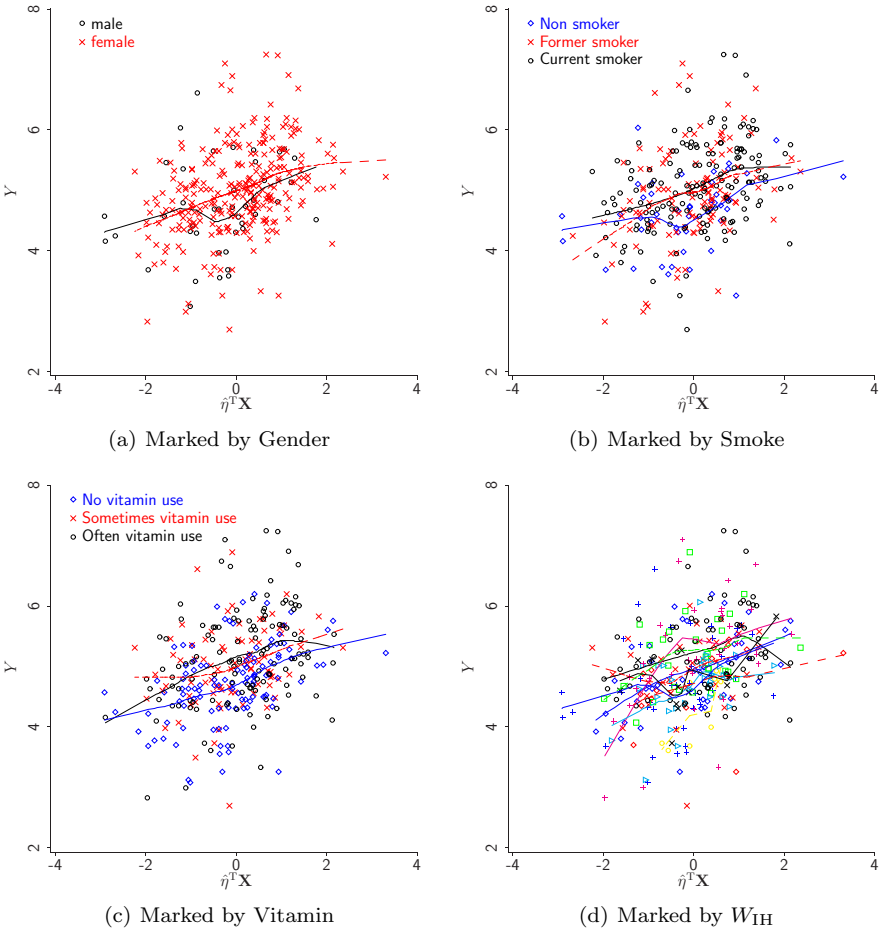
This study was originally done in Nierenberg *et al.* (1989). They found that dietary carotene was positively related to Beta-carotene levels, while Quetelet was negatively related. The data was obtained from *StatLib* webpage and used under permission. Since cases with numbers 257 was suspected as an outlier, they were deleted from the data set, and the total number of sample sizes were 314.

To guarantee the requirements for OPSIR, Calorie, Fiber and Ret.diet were transformed to log-scale and Quetelet to the inverse-scale. In addition, following the suggestion of Nierenberg *et al.* (1989), Beta-carotene plasma concentration was transformed to log-scale for symmetry. After the transformation, we used the following variables for the regression: $Y = \log(\text{Beta} - \text{carotene plasma levels})$; $\mathbf{X} = (\log \text{Calorie}, \log \text{Fiber}, \log \text{Ret.diet}, \text{Quetelet}^{-1})^{\mathrm{T}}$; $\mathbf{W} = (\text{Gender}, \text{Smoke}, \text{Vitamin})^{\mathrm{T}}$.

The data is heavily unbalanced for Gender. The total cases for males are just 42, while there are 272 observations for female. Therefore, instead of W_{FH} for gender, smoke and vitamin, we will consider W_{IH} with 10 levels of 9 cross-combinations between Smoke and Vitamin with females and one level of males.

To investigate that condition 1 or 2 is satisfied, we performed partial dimension reduction with consideration of Gender (G) alone, Smoke (S) alone and Vitamin (V) alone, and define that $\hat{\boldsymbol{\eta}}_{\mathrm{G}}$, $\hat{\boldsymbol{\eta}}_{\mathrm{S}}$ and $\hat{\boldsymbol{\eta}}_{\mathrm{V}}$ are the corresponding estimated basis matrices. Using nominal level 5%, application of OPSIR to the three cases with three slices per each level of Gender, Smoke and Vitamin concluded that one-dimensional linearly transformed predictors were sufficient for each regression. A scatter plot matrix of $\hat{\boldsymbol{\eta}}_{\mathrm{G}}^{\mathrm{T}}\mathbf{X}$, $\hat{\boldsymbol{\eta}}_{\mathrm{S}}^{\mathrm{T}}\mathbf{X}$ and $\hat{\boldsymbol{\eta}}_{\mathrm{V}}^{\mathrm{T}}\mathbf{X}$ is reported in Figure 3.1, which shows that the three have very strong linear relationships. This indicates that either of conditions 1 or 2 seems to hold.

Next we applied OPSIR with two slices per each level of W_{IH}, and determined the structural dimension. The computed p-values for $d = 0$, $d = 1$ and $d = 2$ were 0.001, 0.094 and 0.098



(a) Marked by Gender (b) Marked by Smoke

(c) Marked by Vitamin (d) Marked by W_{IH}

Figure 3.2. Scatter plots of Y and $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ marked by Gender, Smoke, Vitamin and W_{IH} in Section 3.3

respectively. With nominal level 5%, we decide that $\hat{d} = 1$. Then let $\hat{\boldsymbol{\eta}}$, which is shown in Figure 3.1, represent the estimated basis matrix with W_{IH}. Marginally standardizing each of the remaining predictors to have a sample standard deviation of 1, the estimated sufficient predictor $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ is defined:

$$\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X} = -0.687 \log \mathrm{Calorie} + 0.378 \log \mathrm{Fiber} + 0.376 \log \mathrm{Ret.diet} + 0.493 \mathrm{Quetelet}^{-1}.$$

To investigate relations between Y and $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ with Gender, Smoke, Vitamin and W_{IH}, scatter plots were constructed marked by each categorical predictor, and they are reported in Figure 3.2. In Figures 3.2(a)–(d), the colored-lines stands for LOWESS smooths with smoothing parameter 0.7 for each level of each categorical predictor of Gender, Smoke, Vitamin and W_{IH}. According to the figures, there seems no strong interaction between $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ and each of Gender, Smoke, Vitamin and W_{IH}. Therefore, we conclude that additive-linear regression between Y and $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ can summarize the study. Then we can observe the same relation between Beta-carotene plasma concentration levels and dietary factors as Nierenberg *et al.*'s founding.

4. Discussion

Sufficient dimension reduction is restricted to dimension reduction of many-valued or continuous predictors. With a categorical predictor, partial sufficient dimension reduction should be done for many-valued or continuous predictor across subpopulations defined by the categorical predictor. When there are many categorical predictors involved and they are heavily unbalanced, direct application of usual partial dimension reduction methodologies should be problematic and limited in use.

To avoid the issue of unbalanced categorical predictors, we propose usage of an integrated categorical predictor over a hierarchically categorized predictor. To guarantee that the integrated categorical predictor is fully informative to partial dimension reduction, we investigate the required conditions. To see that the conditions hold, one should conduct partial dimension reductions for each categorical predictor, and closely study the correlations of estimated sufficient predictors. Numerical studies confirm its background theory and real data is adequately analyzed through the proposed approach.

Acknowledgements

The authors are grateful to the two referees for many helpful comments.

References

Chiaromonte, F., Cook, R. D. and Li, B. (2002). Dimension reduction with categorical predictors, *Annals of Statistics*, **30**, 475–497.

Cook, R. D. (1998). *Regression Graphics*, Wiley, New York.

Cook, R. D. and Li, B. (2002). Dimension Reduction for the Conditional Mean, *Annals of Statistics*, **30**, 455–474.

Li, B., Cook, R. D. and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors, *Annals of Statistics*, **31**, 1636–1668.

Nierenberg, W. D., Stukel, A. T., Baron, A. J., Dain, J. B. and Greenberg, E. R. (1989). The skin cancer prevention study group, determinants of plasma levels of beta-carotene and retinol, *American Journal of Epidemiology*, **130**, 511–521.

Wen, X. and Cook, R. D. (2007). Optimal sufficient dimension reduction in regressions with categorical predictors, *Journal of Statistical Planning and Inference*, **137**, 1961–1978.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression, *Journal of Royal Statistical Society, Series B*, **64**, 159–175.

Yoo, J. K. (2009). Partial moment-based dimension reduction, *Statistics and Probability Letters*, **79**, 450–456.