ARTICLE IN PRESS

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

# A unified approach to sufficient dimension reduction

Yuan Xue [a,*], Qin Wang [b], Xiangrong Yin [c]

[a] School of Statistics, University of International Business and Economics, Beijing 100029, China
[b] Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA 23284, USA
[c] Department of Statistics, University of Kentucky, Lexington, KY 40508, USA

ARTICLE INFO

*Article history:*
Received 18 April 2017
Received in revised form 9 October 2017
Accepted 2 February 2018
Available online xxxx

*Keywords:*
Dual central subspaces
Hellinger integral
Sufficient dimension reduction
Variable selection

ABSTRACT

Through investigating a recently introduced sufficient dimension reduction method with Hellinger index, this article shows that the generalized Hellinger index unifies three existing dimension reduction methods: kernel discriminant analysis, sliced regression and density minimum average variance estimation, with certain weight functions. The Hellinger index is then extended to regression models with multivariate responses. Furthermore, new algorithms based on Hellinger index to estimate the dual central subspaces and to enable variable selection for sparse models are proposed. Simulation studies and a real data analysis demonstrate the efficacy of the proposed approaches.


## 1. Introduction

Sufficient dimension reduction (SDR) aims at estimating a low-dimensional subspace of a predictor vector $X \in \mathcal{R}^p$ without loss of information on the conditional distribution of the response variable $Y$ given $X$, and without pre-specifying a model for any of $Y|X$ or $X|Y$. This subspace is then called a dimension reduction subspace for the regression of $Y$ on $X$. The intersection of all such subspaces, if itself is a dimension reduction subspace, is called the central subspace (CS) denoted by $\mathcal{S}_{Y|X}$. Its dimension $d_{Y|X}$ is called the structural dimension. We refer readers to Cook (1998a) for more details.

Since *sliced inverse regression* (SIR; Li, 1991) and *sliced average variance estimation* (SAVE; Cook and Weisberg, 1991) were introduced, much attention has been drawn to this area over the past two decades. All proposed methods can be categorized into three groups according to the distribution that is focused on: the inverse regression approach, the forward regression approach and the joint approach. The inverse regression approach focuses on the inverse conditional distribution of $X|Y$. Alongside SIR and SAVE, *principal Hessian directions* (PHD; Li, 1992, 2000; Cook, 1998b), *the kth moment estimation* (Yin and Cook, 2002, 2003), *inverse regression* (Cook and Ni, 2005) and *contour regression* (Li et al., 2005) are well-known methods in this category. They are computationally inexpensive, but require either or both of the linearity and constant covariance conditions (Cook, 1998a). The forward regression approach in which the conditional distribution of $Y|X$ is the object of inference includes *average derivative estimation* (Härdle and Stoker, 1989; Samarov, 1993), *the structure adaptive method* (Hristache et al., 2001), *minimum average variance estimation* (MAVE; Xia et al., 2002), *sliced regression* (SR; Wang and Xia, 2008) and *semiparametric approaches* of Ma and Zhu (2012, 2013a, b, 2014). These methods do not require any strong probabilistic assumptions, however, as either sample size or the number of predictors increases the computational burden increases dramatically due to the use of smoothing. The joint approach includes *Kullback–Leibler distance* (Yin and Cook, 2005; Yin et al., 2008) and *Fourier estimation* (Zhu and Zeng, 2006), which may be flexibly regarded as either inverse or forward approaches.

* Corresponding author.
*E-mail address:* yuanxue@uibe.edu.cn (Y. Xue).

https://doi.org/10.1016/j.jspi.2018.02.001
0378-3758/

Please cite this article in press as: Xue Y., et al., A unified approach to sufficient dimension reduction. J. Statist. Plann. Inference (2018), https://doi.org/10.1016/j.jspi.2018.02.001.

Recently, Wang et al. (2015) proposed a dimension reduction method through the Hellinger integral to estimate the CS. The assumptions are very mild: (a) $\mathcal{S}_{Y|X}$ exists and (b) a finiteness condition, so that the Hellinger integral is always defined. This Hellinger integral approach is computationally efficient and achieves exhaustive estimation of CS. However, Wang et al. (2015) considered only the situation when the response variable is univariate. The innovations of this paper begin with showing that the generalized Hellinger index unifies *kernel discriminant analysis* (Hernández and Velilla, 2005), *sliced regression* (Wang and Xia, 2008) and *density minimum average variance estimation* (Xia, 2007), each being equivalent to adopting suitable weight functions. Then, the Hellinger index is extended to estimate the CS for multivariate regression. We also propose a new algorithm via Hellinger index to estimate the dual central subspaces (DCS), firstly studied by Iaci et al. (2016), where both the CS of $Y|X$ and CS of $X|Y$ are of interests. Finally, the shrinkage estimation procedure proposed by Li and Yin (2008) is incorporated into the Hellinger index in this paper, which enables variable selection and improves the interpretability and accuracy of estimated directions.

The rest of this article is organized as follows. Section 2 briefly reviews the Hellinger index for sufficient dimension reduction. The unification of three existing methods is detailed in Section 3. Section 4 introduces the extension of Hellinger index to estimate CS for multivariate responses. Section 5 describes an efficient approach through Hellinger index to estimate the dual central subspaces. A sparse version of Hellinger method that enables variable selection is discussed in Section 6. Simulated examples and a real data analysis are presented in Sections 7 and 8, respectively. Finally, Section 9 concludes the paper with a short discussion.

## 2. Review on Hellinger index for SDR

Throughout the paper, we assume that the response variable $Y$ and the $p \times 1$ predictor vector $X$ have a joint distribution $F_{(X,Y)}$ with support $\Omega$. Let $\{(x_i,\ y_i), i = 1, \ldots, n\}$ be random samples from $(X, Y)$. The response $Y$ may be continuous or discrete. Traditionally, $X$ is assumed to be continuous in SDR research and we follow this in the paper. The symbol $p(\cdot)$ denotes a probability density function, a probability mass function, or a mixture of the two, whose argument defines implicitly which distribution is referred to. The notation $\int$ is used to denote an integral in the usual sense, a summation, or a mixture of the two. In all cases, $p(x, y) = p(x|y)p(y)$ where $p(x, y)$, $p(x|y)$ and $p(y)$ refer to the joint, conditional and marginal distributions of $(X, Y)$, $X|Y$ and $Y$, respectively.

Let $u, u_1, u_2, \ldots$ denote fixed matrices with $p$ rows and full column rank $d$. The Hellinger integral of order two (Vajda, 1989; Liese and Vajda, 2006) is defined as $H(u) = \mathbb{E}_{X,Y}\left\{R(Y; u^T X)\right\}$, where $R(y; u^T x)$ is the *dependence ratio*, and

$$R(y; u^T x) = \frac{p(y, u^T x)}{p(y)p(u^T x)} = \frac{p(y|u^T x)}{p(y)} = \frac{p(u^T x|y)}{p(u^T x)}. \tag{1}$$

The expectation is over the joint distribution of $X$ and $Y$, which can be emphasized by writing $H(u)$ fully as $H(u; F_{(X,Y)})$.

The joint distribution $F_{(X,Y)}$ for continuous $Y$ or $F_{(X|Y)}$ for discrete $Y$ is assumed to be absolute continuous such that $H(u)$ is finite for all $u$, and thus Hellinger integrals are always defined. Wang et al. (2015) established that $H(u)$ as a natural measure of the regression information $Y|X$ is contained in the space $\mathcal{S} = \text{span}(u)$. Let $\mathcal{H}(\mathcal{S}) = H(u)$, and $\mathcal{S}_1 = \text{span}(u_1)$ and $\mathcal{S}_2 = \text{span}(u_2)$ are two subspaces of $\mathbb{R}^p$ meeting only at the origin, then

$$\mathcal{H}(\mathcal{S}_1 \oplus \mathcal{S}_2) - \mathcal{H}(\mathcal{S}_1) \geq 0,$$

where the equality holds if and only if $Y$ is conditionally independent of $u_2^T X$ given $u_1^T X$.

For a known $d_{Y|X}$, maximizing $H(u)$ over a Grassmann manifold $\mathcal{U}_{d_{Y|X}}$, where $\mathcal{U}_{d_{Y|X}} = \{\text{all } p \times d_{Y|X} \text{ matrices } u \text{ with } u^T u = I_{d_{Y|X}}\}$, recovers a basis of $\mathcal{S}_{Y|X}$. To avoid the curse of dimensionality in estimating the probability densities, Wang et al. (2015) proposed to approximate the dependence ratio with local linearization, and then to aggregate the principal directions to form an estimate of $\mathcal{S}_{Y|X}$.

## 3. Unification of three existing methods

In this section, we show that the Hellinger approach unifies three existing SDR methods, *kernel discriminant analysis* (Hernández and Velilla, 2005), *density minimum average variance estimation* (Xia, 2007) and *sliced regression* (Wang and Xia, 2008). Before presenting the main result, we briefly introduce these three approaches.

Hernández and Velilla (2005) proposed a *kernel discriminant analysis* (KDA) approach in classification setting, where the response variable $Y$ represents the class labels. Suppose $Y$ takes values in a countable index set $\mathcal{Y} \subset \mathbb{R}$, with probability $p(y) = p(Y = y) > 0$ and $\sum_{y\in\mathcal{Y}} p(y) = 1$. The KDA approach is based upon the following functional $T(X) = \sum_{y\in\mathcal{Y}} \text{Var}_X\left\{p(y|X)\right\}$, which measures the aggregate separation among the conditional densities $p(x|Y = y)$ for all $y \in \mathcal{Y}$, where $\text{Var}_X$ denotes the variance with respect to $X$. Clearly, the larger the $T(X)$, the better separation among all classes in $\mathcal{Y}$. When the structural dimension $d_{Y|X} = d$ is known, a basis for the central subspace $\mathcal{S}_{Y|X}$ can be estimated as

$$u^* = \arg\max_{u\in\mathcal{U}_d} T(u^T X) = \arg\max_{u\in\mathcal{U}_d} \sum_{y\in\mathcal{Y}} \text{Var}_X\left\{p(Y = y|u^T X)\right\}, \tag{2}$$

Please cite this article in press as: Xue Y., et al., A unified approach to sufficient dimension reduction. J. Statist. Plann. Inference (2018), https://doi.org/10.1016/j.jspi.2018.02.001.