# Enhancing Sufficient Dimension Reduction via Hellinger Correlation

**Seungbeom Hong** [1] **Ilmun Kim** [2] **Jun Song** [1]

## Abstract

In this work, we develop a new theory and method for sufficient dimension reduction (SDR) in single-index models, where SDR is a subfield of supervised dimension reduction based on conditional independence. Our work is primarily motivated by the recent introduction of the Hellinger correlation as a dependency measure. Utilizing this measure, we develop a method capable of effectively detecting the dimension reduction subspace, complete with theoretical justification. Through extensive numerical experiments, we demonstrate that our proposed method significantly enhances and outperforms existing SDR methods. This improvement is largely attributed to our proposed method's deeper understanding of data dependencies and the refinement of existing SDR techniques.

## 1. Introduction

In the age of big data, the advent of high-dimensional datasets has transformed the landscape of statistical analysis and machine learning. However, as the number of features in a dataset increases, so does the complexity of modeling and interpretation. This phenomenon, often referred to as the curse of dimensionality, poses a significant challenge to researchers and practitioners who seek to extract meaningful information and insights from vast and intricate data structures.

In response to this challenge, the field of sufficient dimension reduction (SDR) has emerged as a powerful approach to navigating high-dimensional space and uncovering the underlying structure without compromising interpretability. Much like how sufficient statistics provide essential information for estimation, sufficient dimension reduction methods furnish us with a subspace that contains adequate information to accurately estimate or explain the response variable. This approach is rooted in the idea that by identifying and preserving the key relationships in the notion of conditional independence as follows. For a univariate random variable $Y$ and a $p$-dimensional random vector $X$, the objective of linear sufficient dimension reduction is to seek out a matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$ $(d < p)$ that possesses the smallest achievable column space such that

$$Y \perp\!\!\!\perp X \,|\, \mathbf{B}^{\mathsf{T}} X. \tag{1}$$

It is important to note that the conditional independence does not change when $\mathbf{B}$ is multiplied by any non-singular matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$. Thus, to make the target identifiable, the parameter needed to seek is the space spanned by the columns of $\mathbf{B}$, i.e, $\mathrm{Span}(\mathbf{B})$, not a matrix $\mathbf{B}$ itself. The column space of $\mathbf{B}$ with the smallest $d$ is called the *central space* denoted by $\mathscr{S}_{Y|X}$ and the dimension of the central space is the structural dimension, say $d$.

The conditional independence (1) can also be represented as

$$Y \perp\!\!\!\perp X \,|\, R(X),$$

for a proper linear mapping, $R : \mathbb{R}^p \to \mathbb{R}^d$. It can be further represented as

$$Y \,|\, X \sim Y \,|\, R(X),$$

where $\sim$ means equal in distribution, which tells us that once $R(X)$ is identified, no more information about $Y$ can be obtained from $X$, and all the regression information for the predictor is preserved through $R(X)$. Another equivalent statement is

$$X \,|\, Y, R(X) \sim X \,|\, R(X).$$

Consider rewriting that $X$ represents data $D$ and $Y$ symbolizes a parameter $\theta$. Under this reinterpretation, the above statement is equivalent to $D \,|\, (\theta, R) \sim D \,|\, R$, which suggests that $R$ acts as a sufficient statistic. Hence, the SDR mapping $R(\cdot)$ aligns with the traditional concept of statistical sufficiency. A key distinction, however, lies in the nature of the sufficient statistic versus the SDR: while a sufficient statistic is observable, the SDR involves parameters that require estimation. Adragni & Cook (2009) explains this conceptual idea of sufficiency more rigorously.

[1]Department of Statistics, Korea University, Seoul, South Korea [2]Department of Applied Statistics, Yonsei University, Seoul, South Korea. Correspondence to: Jun Song <junsong@korea.ac.kr>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

If we have a proper $\mathbf{B}$ satisfying conditional independence as in (1), then the response $Y$ can be represented as

$$Y = g(b_1^\intercal X, \ \ldots, \ b_d^\intercal X, \varepsilon), \tag{2}$$

where $g : \mathbb{R}^{p+1} \mapsto \mathbb{R}$ is an unknown measurable function and $\varepsilon$ is independent of $X$ with mean zero, and $b_1, \cdots, b_d$ are the columns of $\mathbf{B}$. As there is no strong prerequisite for the function $g$, dimension reduction can be performed without relying on a specific model.

Various SDR methods have been proposed and successfully applied in diverse disciplines such as bioinformatics (Chiaromonte & Martinelli, 2002; Hsueh & Tsai, 2016), finance (Wang, 2023), marketing (Naik et al., 2000) and ecology (Roley & Newman, 2008). The two most dominating approaches for SDR are inverse regression and forward regression approach. The inverse method requires an additional assumption on the conditional distribution of the predictors given the response, $X \mid Y$. The well-known techniques for the inverse method include sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook & Weisberg, 1991), and directional regression (DR) (Li & Wang, 2007). The forward method requires less assumptions on the conditional distribution but has additional assumptions on the link function. The minimum average variance estimation (MAVE) (Xia et al., 2002) is a widely recognized approach to the forward method. We refer to Li (2018) for a detailed explanation. Although the prior methods have proven effective, their success relies on specific assumptions mentioned earlier: inverse regression methods impose strong assumptions on the conditional/marginal distribution of $X$ such as the predictors $X$ following an elliptical distribution, while forward-regression approaches require an assumption on the link function $g(\cdot)$. All these assumptions are difficult to verify with the dataset.

On the other hand, there are other directions of SDR studies, which build on measures of statistical dependence. Quantifying the dependence between two random objects has been a central topic in statistics. Some notable examples of dependence measures, especially in nonparametric settings, include the distance covariance (Székely et al., 2007; Székely & Rizzo, 2009), the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2005) and the ball covariance (Pan et al., 2020). The SDR subspace can be found by maximizing the dependence between the response $Y$ and predictors $X$. Sheng & Yin (2013) showed that sufficient dimension reduction can be achieved for a single-index model by the distance covariance. Sheng & Yin (2016) extended this method to a general structural dimension $d$. Similarly, Zhang & Yin (2015) proposed a way to utilize the HSIC for the single index model, and Xue et al. (2017) extended it to a general dimension $d$. In addition, Zhang & Chen (2019) analyzed single and multi-index models based on the ball covariance. See Dong (2021) for a comprehensive review.

While existing SDR methods based on dependence measures have shown their effectiveness, they are not free from limitations. One notable limitation is in the interpretation of the dependence measure itself: while a zero value indicates independence between two random variables, a larger measure does not necessarily imply a stronger relationship between them. This gap can lead to potential misinterpretations when assessing the strength of these relationships. Moreover, these methods encounter challenges regarding their theoretical foundation. Some dependence measure-based SDR methods either lack comprehensive theoretical validation or rely on specific independence assumptions that are difficult to verify in practical applications. It raises concerns about their reliability.

To address these issues, this article introduces a new SDR method using Hellinger correlation, which improves the accuracy of estimating the central subspace while achieving its theoretical justification with weaker assumptions than existing methods. More precisely, compared to other dependence measures, the Hellinger correlation is more adept at capturing the strength of various relationships and satisfies the natural axioms for dependence measures (Geenens & Lafaye de Micheaux, 2022). Equipped with the benefits of the Hellinger correlation, our SDR approach is straightforward to implement and consistently delivers significant improvements over existing methods in almost all cases considered. Moreover, we distinguish the proposed method from existing approaches by establishing a theoretical foundation, including consistency guarantees with minimal assumptions.

To establish a solid foundation for a new SDR method based on a dependence measure, we focus on the single-index model where the target dimension is one. Although extending our method to multi-index models is feasible and promising, as discussed in Section 6 and other work (Sheng & Yin, 2016; Christou, 2020), we have opted to prioritize the foundational principles of SDR by concentrating on the simple yet fundamental single-indexed model. This approach aims to ensure clarity and maintain the robustness of the article, while laying a strong groundwork for future expansions.

The remainder of this article is organized as follows. Section 2 provides background information and motivation for this paper. Section 3 delves into the optimization methods and presents the theoretical results. In Section 4, we provide simulation results that compare our method with existing ones. Section 5 presents a real data application of the method. Section 6 summarizes our contributions and discusses several directions for future work. Lastly, the appendix includes additional simulation results. The code that implements our proposed method is available at https://github.com/JSongLab/SDR_HC.

## 2. Background and Motivation

### 2.1. SDR through dependence measure

As mentioned in Section 1, there have been studies focused on developing SDR methods that utilize dependence measures. To explain the idea, assume that the structural dimension is known as $d$. Let $\boldsymbol{\eta}_0$ be the basis of the central subspace and $\boldsymbol{\eta}_1$ be the basis of the orthogonal complement of the central subspace. In other words, $\boldsymbol{\eta} = (\boldsymbol{\eta}_0, \boldsymbol{\eta}_1)^\top$ is a basis of $\mathbb{R}^p$. Let $\rho : \mathbb{R}^d \times \mathbb{R} \to [0, \infty)$ be a generic dependence measure between two random quantities. A SDR method based on $\rho$ seeks to find the central space $\mathscr{S}_{Y|X}$ by solving the following optimization problem:

$$\mathbf{B}_0 = \underset{\mathbf{B} \in \mathbb{R}^{p \times d}}{\operatorname{argmax}} \rho(\mathbf{B}^\mathsf{T} X, Y) \quad \text{subject to} \quad \mathbf{B}^\mathsf{T} \Sigma_X \mathbf{B} = I_d$$

where $\Sigma_X$ is the covariance matrix of $X \in \mathbb{R}^p$. To conclude that the maximizer recovers the central space, i.e., $\operatorname{Span}(\mathbf{B}_0) = \operatorname{Span}(\boldsymbol{\eta}_0)$, the previous approaches impose an additional independence assumption on projected random variables, namely

$$\boldsymbol{\eta}_0{}^\mathsf{T} X \perp\!\!\!\perp \boldsymbol{\eta}_1{}^\mathsf{T} X,$$

which is not easily verifiable in practical applications. In contrast, we aim to remove this additional restriction and propose a more reliable SDR method.

### 2.2. Copula

We next briefly discuss copulas. Copulas are essential tools in high-dimensional analysis, enabling us to estimate random vectors through the estimation of marginal distributions. This tool has applications across various fields, with finance being a prime example of its extensive usage (Cherubini et al., 2004). Assume that there exists a continuous random vector $X = (X_1, X_2, \ldots, X_p)^\mathsf{T}$. Let $U_i = F_i(X_i)$ where $F_i$ is the cumulative distribution function of $X_i$. By the integral probability transform, all $U_i$ are uniform random variables on the interval $[0, 1]$. The copula $C$ is defined as

$$C(u_1, \ldots, u_p) = P(U_1 \le u_1, \ldots, U_p \le u_p). \quad (3)$$

In other words, the copula $C$ is the joint cumulative distribution function of a random vector in the unit cube, where each marginal is a uniform random variable.

Sklar (1959) explains that a copula is an adequate tool for understanding the distribution of a random vector, and establishes the following result:

**Theorem 2.1** (Sklar, 1959). *Let $X = (X_1, X_2, \ldots, X_p)^\mathsf{T}$ be a random vector. Suppose that $F$ and $f$ are the joint cumulative distribution function and the joint probability density function of $X$. Then, there exists a function $C : [0, 1]^p \to [0, 1]$, called the copula of $X$, such that*

$$F(x_1, x_2, \ldots, x_p) = C(F_1(x_1), F_2(x_2, ), \ldots, F_p(x_p)).$$

*Additionally, there exists a function $c : [0, 1]^p \to [0, \infty)$, called the copula density of $X$, such that*

$$f(x_1, x_2, \ldots, x_p) = c(F_1(x_1), F_2(x_2, ), \ldots, F_p(x_p)) \times f_1(x_1) f_2(x_2) \cdots f_p(x_p).$$

For a random variable $Y$, we have $F_Y(y) = C(F_Y(y))$ and $f_Y(y) = c(F_Y(y)) f_Y(y)$. Thus, the univariate random variable copula is the identity function and the density of the copula is 1 in the unit interval [0, 1]. This property will be used later in the construction of the proposed method.

Since the cumulative distribution function of a random variable is a monotonic function, a copula is invariant to the monotonic transformation of its marginals. This property is powerful in determining the dependence between random variables.

### 2.3. f-divergence

The $f$-divergence is a function that measures the difference between two distributions $P$ and $Q$ given as

$$D_\varphi(P \| Q) = \int \varphi\left(\frac{dP}{dQ}\right) dQ, \quad (4)$$

where $\varphi : (0, \infty) \to \mathbb{R}$ is convex and $\varphi(1) = 0$. The $f$-divergence family encompasses a wide range of statistical distances between distributions. Some notable examples include the Kullback–Leibler divergence with $\varphi(t) = t \log t$, the squared Hellinger distance with $\varphi(t) = (\sqrt{t} - 1)^2$ and the total variation distance with $\varphi(t) = |t - 1|/2$.

If $\varphi$ is strictly convex, $P$ and $Q$ are identical distributions if and only if $D_\varphi(P \| Q) = 0$. This characteristic property allows us to build upon the $f$-divergence to test the independence between two random vectors. More formally, the hypotheses for independence testing are given as

$$H_0 : F_{XY} = F_X F_Y \quad \text{versus} \quad H_1 : F_{XY} \neq F_X F_Y,$$

where $F_{XY}$ is the joint distribution of $(X, Y)$ and $F_X F_Y$ is the product of the marginal distributions. The characteristic property of the $f$-divergence implies that the above hypotheses can be equivalently written as $H_0 : D_\varphi(F_{XY} \| F_X F_Y) = 0$ versus $H_1 : D_\varphi(F_{XY} \| F_X F_Y) \neq 0$. Hence one can use an estimator of

$$D_\varphi(F_{XY} \| F_X F_Y) = \iint \varphi\left(\frac{dF_{XY}}{dF_X dF_Y}\right) dF_X dF_Y$$

as a test statistic for independence testing. If both $F_X$ and $F_Y$ are absolutely continuous, then $D_\varphi$ can be represented in terms of density functions as follows:

$$D_\varphi(F_{XY} \| F_X F_Y) = \iint f_X f_Y \varphi\left(\frac{f_{XY}}{f_X f_Y}\right) dx dy. \quad (5)$$

Kinney & Atwal (2014) proved that any measure of dependence constructed by $f$-divergence (5) holds data processing inequality. That is, $D(X,Z) \leq D(Y,Z)$ where $X \perp\!\!\!\perp Z \,|\, Y$.

### 2.4. Hellinger Correlation

Geenens & Lafaye de Micheaux (2022) proposed the Hellinger correlation as a tool for capturing the dependence between a pair of random variables. As implied by its name, it is formulated based on the squared Hellinger distance, which is an example of the $f$-divergence. More specifically, the squared Hellinger distance between $F_{XY}$ and $F_X F_Y$ is given as

$$\begin{aligned}\mathcal{H}^2(X,Y) &= \frac{1}{2}\iint_{\mathbb{R}^2}\left(\sqrt{\frac{dF_{XY}}{dF_X dF_Y}}-1\right)^2 dF_X dF_Y \\ &= \frac{1}{2}\iint_{\mathcal{I}^2}(\sqrt{c_{XY}(u_x,u_y)}-1)^2 du_x du_y \\ &= 1-\iint_{\mathcal{I}^2}\sqrt{c_{XY}(u_x,u_y)}du_x du_y \\ &:= 1-\mathcal{B}(X,Y),\end{aligned} \tag{6}$$

where $\mathcal{I}^2$ denotes the unit square $[0,1]^2$. In the above equations, $c_{XY}$ denotes the joint copula density of $U_X$ and $U_Y$ where $U_X$ and $U_Y$ denote the cumulative distribution function of $X$ and $Y$, respectively. The quantity $\mathcal{B}(X,Y)$ in the last line is referred to as the Bhattacharyya affinity coefficient (Bhattacharyya, 1943) between the copula densities. From now on, we will write $\mathcal{B}(X,Y)$ as $\mathcal{B}$ for simplicity.

To motivate the Hellinger correlation, consider a bivariate normal random vector $(X,Y) \sim N((0,0),\left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right))$. As dicsussed in Geenens & Lafaye de Micheaux (2022), the squared Hellinger distance between $X$ and $Y$ has an explicit expression as

$$\mathcal{H}^2(X,Y) = 1-(2(1-\rho^2)^{1/4})/(4-\rho^2)^{1/2}$$

or $\mathcal{B} = (2(1-\rho^2)^{1/4})/(4-\rho^2)^{1/2}$ in terms of the Bhattacharyya affinity coefficient. As a result, the correlation parameter $\rho$ can be written as

$$|\rho| = \frac{2}{\mathcal{B}^2}\{\mathcal{B}^4+(4-3\mathcal{B}^4)^{1/2}-2\}^{1/2}.$$

This relationship leads to the Hellinger correlation between random variables $X$ and $Y$ defined as follows.

**Definition 2.2.** Let $\mathcal{B}$ denote the Bhattacharyya affinity coefficient for $(X,Y)$ defined in (6). The Hellinger correlation between $X$ and $Y$ is defined as

$$H(X,Y) = \frac{2}{\mathcal{B}^2}\{\mathcal{B}^4+(4-3\mathcal{B}^4)^{1/2}-2\}^{1/2}. \tag{7}$$

By construction, the Hellinger correlation $H(X,Y)$ coincides with the Pearson correlation when $(X,Y)$ follows a joint normal distribution, whereas they can differ significantly for non-normal distributions. It is worth noting that the properties of $f$-divergence and copula are preserved in the Hellinger correlation as the map $h : [0,1] \to [0,1]$, given as $h(x) = 2x^{-2}\{x^4+(4-3x^4)^{1/2}-2\}^{1/2}$, is a bijection.

The Hellinger correlation has several attractive properties, worth highlighting. First of all, like distance correlation, the Hellinger correlation fully characterizes independence, i.e., $H(X,Y)=0$ if and only if $X$ and $Y$ are independent. However, unlike distance correlation, the Hellinger correlation does not depend on any moment conditions. Moreover, it is properly normalized as $0 \leq H(X,Y) \leq 1$, and equals one when $X$ and $Y$ are deterministically predictable from each other. See (P6) in Geenens & Lafaye de Micheaux (2022) for a more precise statement. The latter property is in sharp contrast to other popular measures such as Pearson's correlation, distance correlation and rank-based correlations. In particular, Pearson's correlation and distance correlation are 1 if a random variable is an affine transformation of the other variable. Additionally, rank-based measures such as Spearman's $\rho$, Kendall's $\tau$, and Hoeffeding's $D$ are 1 if two random variables have a monotonic deterministic relationship. More fundamentally, the Hellinger correlation takes 1 if and only if there exists a Borel function $\Phi : [0,1] \to \mathbb{R}^2$ such that $(X,Y) = \Phi(U)$ where $U$ is a uniform random variable in the interval $[0,1]$. Another important property of the Hellinger correlation is that it is invariant to any monotonic transformations. This means that for any two strictly monotonic functions $\psi_1, \psi_2$, the following relationship holds

$$H(\psi_1(X),\psi_2(Y)) = H(X,Y).$$

This invariance property has been highlighted as a fundamental property of any valid dependence measure. We refer the reader to Geenens & Lafaye de Micheaux (2022) for further discussion on the properties of the Hellinger correlation. We also point out that the Hellinger correlation tends to be more sensitive to non-linear and realistic dependence than other popular dependence measures as illustrated in the simulation section in Geenens & Lafaye de Micheaux (2022).

Definition 2.2 indicates that estimating $\mathcal{B}$ is sufficient for estimating $H(X,Y)$. Geenens & Lafaye de Micheaux (2022) introduce an estimator of $\mathcal{B}$ based on the estimator of Leonenko et al. (2008). To explain, let $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ be a random sample of size $n$. Let $U_i = (U_{X_i},U_{Y_i}) = (F_X(X_i),F_Y(Y_i))$. Under the continuity assumption for $F_X$ and $F_Y$, it is clear that $U_{X_i}$ and $U_{Y_i}$ are uniform random variables. Let $\hat{U}_i$, the sample version of $U_i$, be $(\hat{F}_X(X_i),\hat{F}_Y(Y_i))$ where

$\hat{F}_X(u) := (1/(n+1))\sum_{i=1}^n I_{\{X_i \le u\}}$ and $\hat{F}_Y(u) := (1/(n+1))\sum_{i=1}^n I_{\{Y_i \le u\}}$. Let $R_i = \min_{j\neq i} \|U_j - U_i\|_2$ and $\hat{R}_i = \min_{j\neq i} \|\hat{U}_j - \hat{U}_i\|_2$. Then the final estimator of $\mathcal{B}$ suggested by Geenens & Lafaye de Micheaux (2022) is

$$\hat{\mathcal{B}}_n = \frac{2\sqrt{n-1}}{n}\sum_{i=1}^n \hat{R}_i, \tag{8}$$

and the corresponding estimator of the Hellinger correlation is

$$\hat{H}_n(X,Y) = \frac{2}{\hat{\mathcal{B}}_n^2}\{\hat{\mathcal{B}}_n^4 + (4-3\hat{\mathcal{B}}_n^4)^{1/2} - 2\}^{1/2}. \tag{9}$$

Our work builds upon these estimators of the Bhattacharyya affinity coefficient and the Hellinger correlation, and proposes an SDR method that offers both theoretical and empirical advantages over existing approaches.

There are several instances of $f$-divergence (e.g., the total variation distance) that share similar properties as the Hellinger correlation. However, unlike the Hellinger correlation, there is currently a lack of computationally efficient estimators with solid theoretical guarantees for these divergences, which is the main bottleneck for using those in SDR applications. We therefore focus on the Hellinger correlation in this work, while leaving the exploration of other $f$-divergence measures for SDR as an interesting avenue for future research.

## 3. Main Results

We now introduce the main results of this work by focusing on the setting where the structural dimension is $d = 1$ and both $X$ and $Y$ have continuous distributions. In this setting, the multi-index model (2) becomes the single-index model

$$Y = g(\eta_0{}^\mathsf{T} X, \varepsilon), \tag{10}$$

and our goal is then to estimate the central subspace spanned by $\eta_0$ through the Hellinger correlation.

### 3.1. Method

For the purpose of identification, we restrict the parameter space to the unit sphere of $\mathbb{R}^p$, which is denoted as $\mathbb{S}^{p-1}$. Since $H$ is a monotonically decreasing function of $\mathcal{B}$, minimizing $\mathcal{B}$ is equivalent to maximizing $H$. Thus, our objective is to find $\eta_0$ such that

$$\eta_0 = \operatorname*{argmax}_{\alpha\in\mathbb{S}^{p-1}} H(\alpha^\mathsf{T} X, Y) = \operatorname*{argmin}_{\alpha\in\mathbb{S}^{p-1}} \mathcal{B}(\alpha^\mathsf{T} X, Y). \tag{11}$$

We use the sphere coordinate to represent $\mathbb{S}^{p-1}$. To represent the direction vector in the Euclidean coordinate, we convert it to the $p-1$ radian tuple. For $\alpha \in \mathbb{R}^p$, there exists $\phi = (\phi_1, \phi_2, \ldots, \phi_{p-1})$ where $\phi_1, \ldots, \phi_{p-2} \in [0, \pi]$ and $\phi_{p-1} \in [0, 2\pi)$ defined as below:

$$\begin{aligned}
\phi_1 &= \arctan\left(\sqrt{\alpha_p^2 + \cdots + \alpha_2^2}/\alpha_1\right) \\
\phi_2 &= \arctan\left(\sqrt{\alpha_p^2 + \cdots + \alpha_3^2}/\alpha_2\right) \\
&\vdots \\
\phi_{p-1} &= \arctan(\alpha_p/\alpha_{p-1}).
\end{aligned} \tag{12}$$

Given the radian tuple, our optimization process consists of two steps. First, we use the simulated annealing method (Bélisle, 1992) given initial values produced by existing SDR methods: SIR, SAVE, DR, and MAVE. Second, starting with the results of the first method, we apply the downhill simplex method proposed by Nelder & Mead (1965). After optimization, we transform $\phi$ and return $(\alpha_1, \ldots, \alpha_p)^\top \in \mathbb{R}^p$ defined as

$$\begin{aligned}
\alpha_1 &= \cos(\phi_1) \\
\alpha_2 &= \sin(\phi_1)\cos(\phi_2) \\
&\vdots \\
\alpha_{p-1} &= \sin(\phi_1)\cdots\sin(\phi_{p-2})\cos(\phi_{p-1}) \\
\alpha_p &= \sin(\phi_1)\cdots\sin(\phi_{p-2})\sin(\phi_{p-1}).
\end{aligned} \tag{13}$$

The sample-level estimation procedure is based on our estimators for $\mathcal{B}$ and $H$ given in (8) and (9):

$$\hat{\eta}_n = \operatorname*{argmax}_{\alpha\in\mathbb{S}^{p-1}} \hat{H}_n(\alpha^\mathsf{T} X, Y) = \operatorname*{argmin}_{\alpha\in\mathbb{S}^{p-1}} \hat{\mathcal{B}}_n(\alpha^\mathsf{T} X, Y),$$

and our next goal is to investigate theoretical and empirical properties of $\hat{\eta}_n$.

Before moving on, let us briefly discuss the computational complexity of the proposed procedure. Our method involves computing the Hellinger correlation estimator, which has a complexity of $O(n^2 p)$. The transformation of $\alpha$ into spherical coordinates and back into Euclidean coordinates adds a complexity of $O(p)$ per iteration, maintaining the overall complexity at $O(n^2 p)$ per iteration. The number of iterations for the downhill simplex method to reach a local optimum varies depending on several factors such as initial values and tolerance, making precise complexity analysis challenging. Nevertheless, denoting the number of iterations as $k$, the overall complexity of our method can be concisely written as $O(n^2 pk)$.

### 3.2. Theoretical Results

In this section, we show the consistency of the sample-level estimation $\hat{\eta}_n$. First, we show that the population-level estimation (11) recovers the central space and the solution is unique up to a sign-flip. Specifically, the next

theorem shows that we can recover the central subspace by maximizing $H(\alpha^\mathsf{T}X, Y)$ with respect to $\alpha \in \mathbb{S}^{p-1}$, i.e., $\eta_0 = \operatorname{argmax} H(\alpha^\mathsf{T}X, Y)$ over all $\alpha \in \mathbb{S}^{p-1}$.

**Theorem 3.1.** *Let $X \in \mathbb{R}^p$ be a random vector and $Y \in \mathbb{R}$ be a random variable. Let $\eta_0 \in \mathbb{S}^{p-1}$ be the basis of the central subspace $\mathscr{S}_{Y|X}$. Then $\eta_0 = \operatorname{argmax} H(\alpha^\mathsf{T}X, Y)$ for all $\alpha \in \mathbb{S}^{p-1}$. Moreover it holds that $H(\eta_0{}^\mathsf{T}X, Y) = H(\alpha^\mathsf{T}X, Y)$ if and only if $\operatorname{Span}(\eta_0) = \operatorname{Span}(\alpha)$.*

*Proof.* Geenens & Lafaye de Micheaux (2022) explained that the Hellinger correlation satisfies the generalized data processing inequality. That is, $H(X, Y) \leq \min\{H(X, Z), H(Y, Z)\}$ if $X \perp\!\!\!\perp Y \mid Z$. Let $P_\alpha$ be the orthogonal projection matrix generated by $\alpha$. In addition, let $\sigma(A)$ denote the smallest $\sigma$-algebra generated by the random variable $A$. From sufficient dimension reduction assumptions, for any $\alpha \in \mathbb{S}^{p-1}$,

$$X \perp\!\!\!\perp Y \mid \eta_0^\mathsf{T}X \Rightarrow P_\alpha X \perp\!\!\!\perp Y \mid \eta_0{}^\mathsf{T}X$$

since $\sigma(P_\alpha X) \subseteq \sigma(X)$. We also have $\sigma(P_\alpha \mathbf{x}) = \sigma(\alpha^\mathsf{T}\mathbf{x})$. Thus,

$$\alpha^\mathsf{T}X \perp\!\!\!\perp Y \mid \eta_0{}^\mathsf{T}X.$$

By the property of the Hellinger correlation,

$$H(\alpha^\mathsf{T}X, Y) \leq H(\eta_0{}^\mathsf{T}X, Y).$$

Thus, $H(\alpha^\mathsf{T}X, Y)$ achieves the maximum when $\alpha = \eta_0$.

Next we prove that $H(\eta_0^\mathsf{T}X, Y) = H(\alpha^\mathsf{T}X, Y)$ *if and only if* $\operatorname{Span}(\eta_0) = \operatorname{Span}(\alpha)$. The "*if*" direction is trivial because the Hellinger correlation is invariant to monotonic transformations. We thus focus on the "*only if*" direction.

Suppose now that $\alpha_0 \in \mathbb{S}^{p-1}$ is another maximizer. If there is a monotonic relation between $\alpha_0{}^\mathsf{T}X$ and $\eta_0{}^\mathsf{T}X$, then $\sigma(\alpha_0{}^\mathsf{T}X) = \sigma(\eta_0{}^\mathsf{T}X)$ and $\alpha_0 \in \mathscr{S}_{Y|X}$. We next assume that there is no monotonic relationship between $\alpha_0{}^\mathsf{T}X$ and $\eta_0{}^\mathsf{T}X$, and show that this will contradict our condition that $\alpha_0$ is another maximizer of $H(\alpha^\top X, Y)$. Since the Hellinger correlation is a monotone increasing function of the squared Hellinger distance, it suffices to prove that $\mathcal{H}^2(X_1, Y) < \mathcal{H}^2(X_2, Y)$ where $X_1 = \alpha_0{}^\mathsf{T}X$ and $X_2 = \eta_0{}^\mathsf{T}X$. Since $X_1$ and $X_2$ do not have a monotonic relationship, the density functions are written as $p(x_1, y) = \int p(x_1|x_2)p(x_2, y)dx_2$ and $p(x_1) = \int p(x_1|x_2)p(x_2)dx_2$. Equipped with this notation, we have

$$\begin{aligned}
&\mathcal{H}^2(X_1, Y) \\
&= \iint \varphi\left(\frac{p(x_1, y)}{p(x_1)p(y)}\right) p(x_1)p(y)dx_1dy \\
&= \iiint \varphi\left(\frac{\int p(x_1|x_2)p(x_2, y)dx_2}{\int p(x_1|x_2)p(x_2)p(y)dx_2}\right) \\
&\qquad \times p(x_1|x_2)p(x_2)p(y)dx_1dx_2dy \\
&\leq \iiint \varphi\left(\frac{p(x_2, y)}{p(x_2)p(y)}\right) p(x_1|x_2)p(x_2)p(y)dx_1dx_2dy, \\
&= \iint \varphi\left(\frac{p(x_2, y)}{p(x_2)p(y)}\right) p(x_2)p(y)dx_2dy \\
&= \mathcal{H}^2(X_2, Y),
\end{aligned}$$

where the inequality in the third line comes from Jensen's inequality as used in the proof of Theorem 4 in the appendix of Kinney & Atwal (2014). Notice that the the squared Hellinger distance $\mathcal{H}^2$ uses $\varphi(t) = (t^{1/2} - 1)^2$, which is strongly convex. Thus, the equality holds if and only if $\frac{p(x_2,y)}{p(x_2)p(y)} = 1$. In other words, the equality holds if and only if $\eta_0{}^\mathsf{T}X \perp\!\!\!\perp Y$, which contradicts the assumption that $\eta_0 \in \mathbb{S}^{p-1}$ is the basis of the central subspace $\mathscr{S}_{Y|X}$. Therefore, $\alpha_0{}^\mathsf{T}X$ and $\eta_0{}^\mathsf{T}X$ have a monotonic relationship and $\operatorname{Span}(\eta_0) = \operatorname{Span}(\alpha_0)$.

□

Theorem 3.1 indicates that one can find a basis of the central subspace by optimizing the Hellinger correlation or Bhattacharyya affinity coefficient. Since $\eta_0$ and $-\eta_0$ span the same space, the result of optimization is not unique. However, the parameter that one wants to obtain is the central space. Thus, our goal is identifiable.

The next theorem shows that, for any direction vector, the sample-level Hellinger correlation of our objective function is consistent.

**Theorem 3.2.** *Let $\alpha \in \mathbb{S}^{p-1}$ be an arbitrary vector. Then*

$$\hat{H}_n(\alpha^\mathsf{T}X, Y) \xrightarrow{P} H(\alpha^\mathsf{T}X, Y) \tag{14}$$

*where $\xrightarrow{P}$ means convergence in probability.*

*Proof.* Since $H(\alpha^\mathsf{T}X, Y)$ is a continuous function of $\mathcal{B}(\alpha^\mathsf{T}X, Y)$, it suffices to show that $\hat{\mathcal{B}}_n$ converges to $\mathcal{B}$. Let $U = (U_{\alpha^\mathsf{T}X}, U_Y)$ for which it holds that $\mathcal{B} = \mathrm{E}[c^{-1/2}(U)]$. Since $\hat{U}_i$ converges to $U_i$ in probability, $\hat{R}_i$ also converges to $R_i$. Then

$$\hat{\mathcal{B}}_n - \tilde{\mathcal{B}}_n \xrightarrow{P} 0, \tag{15}$$

where $\tilde{\mathcal{B}}_n = \frac{2\sqrt{n-1}}{n} \sum_{i=1}^n R_i$.

Leonenko et al. (2008) provides an estimator for $\mathrm{E}[f(X)^q]$ where $f$ is the density function of $X$. Our estimator $\hat{\mathcal{B}}_n$ corresponds to the case where $q = -1/2$. Theorem 3.2 of Leonenko et al. (2008) shows that $\tilde{\mathcal{B}}_n$ converges to $\mathcal{B}$ in probability. Thus,

$$\hat{\mathcal{B}}_n - \mathcal{B} = (\hat{\mathcal{B}}_n - \tilde{\mathcal{B}}_n) + (\tilde{\mathcal{B}}_n - \mathcal{B}) \xrightarrow{P} 0. \qquad \square$$

With the two above theorems, we can show that the sample-level estimator $\hat{\eta}_n$ converges to the true SDR direction in probability.

**Theorem 3.3.** *Let* $\hat{\eta}_n = \operatorname{argmax}\{\hat{H}_n(\alpha^\intercal X, Y) \,|\, \alpha \in \mathbb{S}^{p-1}\}$ *and* $\eta_0 = \mathbb{S}^{p-1}$ *be the basis for the central subspace* $\mathscr{S}_{Y|X}$. *Then* $z\hat{\eta}_n \xrightarrow{P} \eta_0$ *where* $|z| = 1$.

*Proof.* Suppose that $\hat{\eta}_n$ is not a consistent estimator of $\eta_0$. Since $\mathbb{S}^{p-1}$ is a compact set, $\{\hat{\eta}_n\}$ has a subsequence $\{\hat{\eta}_{m(n)}\}$ that converges to $\eta_*$ where $\operatorname{Span}(\eta_*) \neq \operatorname{Span}(\eta_0)$. Then $\hat{H}_{m(n)}(\hat{\eta}_{m(n)}^\intercal X, Y) \geq \hat{H}_{m(n)}(\eta_0{}^\intercal X, Y)$. If we take a limit on both sides, we obtain

$$H(\eta_*{}^\intercal X, Y) \geq H(\eta_0{}^\intercal X, Y).$$

By Theorem 3.1, there is a contradiction since $\eta_0 = \operatorname{argmax} H(\alpha^\intercal X, Y)$. Thus, $\hat{\eta}_n$ is a consistent estimator of $\eta_0$. □

Like the population-level approach, maximizing the estimate of the Hellinger correlation may give two different results, $\hat{\eta}_n$ or $-\hat{\eta}_n$. The role of $z$ is to equalize the direction. If we focus on the projection matrix, we can check that two direction vectors align with the same subspace.

We emphasize that our results are derived without the assumption $\eta_0{}^\top X \perp\!\!\!\perp \eta_1{}^\top X$ where $\eta_0{}^\top \eta_1 = 0$, which is required for other SDR methods based on dependence measures. This assumption may hold asymptotically for distributions satisfying certain moment conditions. However, in the case of long-tailed distributions, such as the Cauchy distribution, this assumption may not be valid. For further details, see Diaconis & Freedman (1984).

## 4. Numerical Experiments

To evaluate the accuracy of our proposed method, we conducted simulation experiments under various scenarios:

$$\begin{aligned}\text{Model I}: Y &= (\eta^\top X)^2 + \varepsilon.\\ \text{Model II}: Y &= \exp(\eta^\top X) + \varepsilon.\\ \text{Model III}: Y &= 5\sin(\eta^\top X) + \varepsilon,\end{aligned}$$

where $\varepsilon \sim N(\mu = 0, \sigma = 0.2)$ and we set $\eta$ as

$$\begin{aligned}\text{Model I}: \eta &= (1,-1,0,0,0,0,0,0,0,0)^\intercal.\\ \text{Model II}: \eta &= (1,1,1,1,1,0,0,0,0,0)^\intercal.\\ \text{Model III}: \eta &= (1,1,0,0,0,0,0,0,0,0)^\intercal.\end{aligned}$$

We generate $X$ from two different distributions described as

$$\begin{aligned}\text{Normal: } & (X_1,\ldots,X_{10})^\intercal \sim N_{10}(0, I_{10})\\ \text{Non-normal: } & X_1 \sim \text{Exp}(2), X_2 \sim \text{Exp}(4), X_3 \sim \chi^2(5),\\ & X_4 \sim t(15), X_5 \sim t(3),\\ & (X_6,\ldots,X_{10})^\intercal \sim N_5(0, I_5).\end{aligned}$$

To assess the performance of our method, we employ the following metric to quantify the difference between two subspaces:

$$\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated}) = \|P_{\mathcal{S}_{True}} - P_{\mathcal{S}_{Estimated}}\|, \quad (16)$$

where $\|\cdot\|$ is the maximum eigenvalue of a matrix and $P_{\mathcal{S}_{True}}$ and $P_{\mathcal{S}_{Estimated}}$ are the orthogonal projection matrices of the subspace $\mathcal{S}_{True} = \operatorname{Span}(\eta_*)$ and $\mathcal{S}_{Estimated} = \operatorname{Span}(\hat{\eta})$. A smaller value of $\Delta$ indicates a more accurate estimation.

In addition, to provide a robust comparison of the methods, we generate 100 samples of each case with different sample sizes $n = 100, 200, 400$. Then we compute the mean and standard deviation of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples and summarize them in Table 1–Table 3 and also in the appendix.

Table 1 shows the results of the experiment under Model I with the normal predictors. It presents $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ when we use SIR (Li, 1991), SAVE (Cook & Weisberg, 1991), DR (Li & Wang, 2007), and MAVE (Xia et al., 2002). Then SIR-HC, SAVE-HC, DR-HC, MAVE-HC are our methods with their initial values in the iteration as SIR, SAVE, DR, and MAVE, respectively. One can check that SIR fails to recover the central subspace because of the U-shape structure in the model. However, our proposed method based on SIR successfully discloses the central subspace. This result shows that our proposed method can overcome the weakness of initial methods. Furthermore, our method enhances the SDR performance significantly even with the MAVE, which is known to be a gold standard. More importantly, the accuracy increases as $n$ increases. It shows an experimental justification of the consistency of our method.

Table 2 shows the results of the experiment under Model II with the normal predictors. Model II has a strong monotonic relation in which SAVE does not perform well. Similar to Model I, our method can capture central space effectively even with the worst case, and improves its accuracy significantly.

Table 3 presents the summary of results for Model III with the normal predictors. We can see that the inverse-regression methods such as SIR, SAVE, and DR require a larger sample size to capture the direction correctly. However, with our proposed method, all the estimators become closer to the true direction with high accuracy even with the small sample size.

Figure 1 provides the boxplots of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ for Model I, II, and II with $n = 100$ and normal predictors. Overall, our method improves existing SDR methods effectively in various scenarios.

Additional comparisons with contemporary SDR methods

*Table 1.* Model I: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when predictors are normal.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| $n = 100$ | 0.9546 | 0.037 | 0.5154 | 0.0355 | 0.2995 | 0.0371 | 0.0654 | 0.0379 |
| | (0.0801) | (0.0211) | (0.1712) | (0.0194) | (0.0744) | (0.0202) | (0.0223) | (0.0203) |
| $n = 200$ | 0.8868 | 0.0261 | 0.2977 | 0.0242 | 0.1961 | 0.0232 | 0.0353 | 0.0257 |
| | (0.1552) | (0.014) | (0.098) | (0.0136) | (0.0534) | (0.0126) | (0.0087) | (0.0109) |
| $n = 400$ | 0.8793 | 0.0183 | 0.1938 | 0.0177 | 0.1262 | 0.0183 | 0.0205 | 0.0195 |
| | (0.1603) | (0.0082) | (0.0485) | (0.0096) | (0.0278) | (0.009) | (0.0053) | (0.0095) |

*Table 2.* Model II: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when predictors are normal.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| $n = 100$ | 0.1191 | 0.0737 | 0.9937 | 0.1203 | 0.1659 | 0.0732 | 0.0699 | 0.0602 |
| | (0.0365) | (0.0406) | (0.0078) | (0.0836) | (0.0475) | (0.0452) | (0.0209) | (0.0251) |
| $n = 200$ | 0.0747 | 0.0356 | 0.7806 | 0.0654 | 0.1007 | 0.0328 | 0.0405 | 0.0314 |
| | (0.0203) | (0.0182) | (0.2891) | (0.0576) | (0.0323) | (0.0144) | (0.0104) | (0.0119) |
| $n = 400$ | 0.0535 | 0.0209 | 0.0691 | 0.023 | 0.0662 | 0.0208 | 0.0261 | 0.0179 |
| | (0.0152) | (0.0092) | (0.0231) | (0.0132) | (0.0161) | (0.0093) | (0.0067) | (0.0071) |

*Table 3.* Model III: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when predictors are normal.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| $n = 100$ | 0.2873 | 0.0186 | 0.9663 | 0.0183 | 0.4165 | 0.0172 | 0.0455 | 0.018 |
| | (0.1141) | (0.0102) | (0.06) | (0.0094) | (0.1897) | (0.0089) | (0.0149) | (0.0095) |
| $n = 200$ | 0.2011 | 0.0135 | 0.9131 | 0.0135 | 0.2606 | 0.0127 | 0.0218 | 0.0144 |
| | (0.0583) | (0.0069) | (0.1254) | (0.0072) | (0.0919) | (0.007) | (0.0063) | (0.0073) |
| $n = 400$ | 0.1433 | 0.0092 | 0.4004 | 0.0096 | 0.1761 | 0.0085 | 0.012 | 0.0094 |
| | (0.0372) | (0.0053) | (0.2249) | (0.0047) | (0.0513) | (0.0042) | (0.0026) | (0.0046) |

using distance covariance, HSIC, and ball covariance are provided in the appendix, along with simulation results with non-sparse $\eta$ and non-normal predictors.

## 5. Real Data Analysis

We apply our methods to the real estate valuation dataset in the UCI Machine Learning Repository (Yeh, 2018). There are 414 observations, and the features in the dataset are

- Transaction date,
- house age,
- distance to the nearest MRT station,
- number of convenience stores,
- latitude,
- longitude, and
- (Target) house price of unit area.

We remove the transaction date variable and standardize the predictors before applying SDR methods. Subsequently, we randomly divide the dataset into a training sample of size 300 and use the remaining objects as the test sample. SDR methods are then applied to the training set to extract the SDR direction, followed by fitting a local polynomial regression using the remaining variables to predict the house price. The weights are given equally for each observation and quadratic polynomial was used to fit the model. Finally, we predict the house price in the test set and compute the test MSE to evaluate the SDR performance. It is important to note that in real data applications, the true central space is unknown, which is why we apply local polynomial regression between the target and the reduced predictor $\hat{\eta}^{\mathsf{T}} X$ to measure the performance. The results are summarized in Table 4.

Table 4 shows that all the SDR methods have been improved with our method. The estimated direction with MAVE-HC is

$$\hat{\eta} = (-0.089, -0.987, 0.019, 0.067, 0.117)^{\mathsf{T}},$$

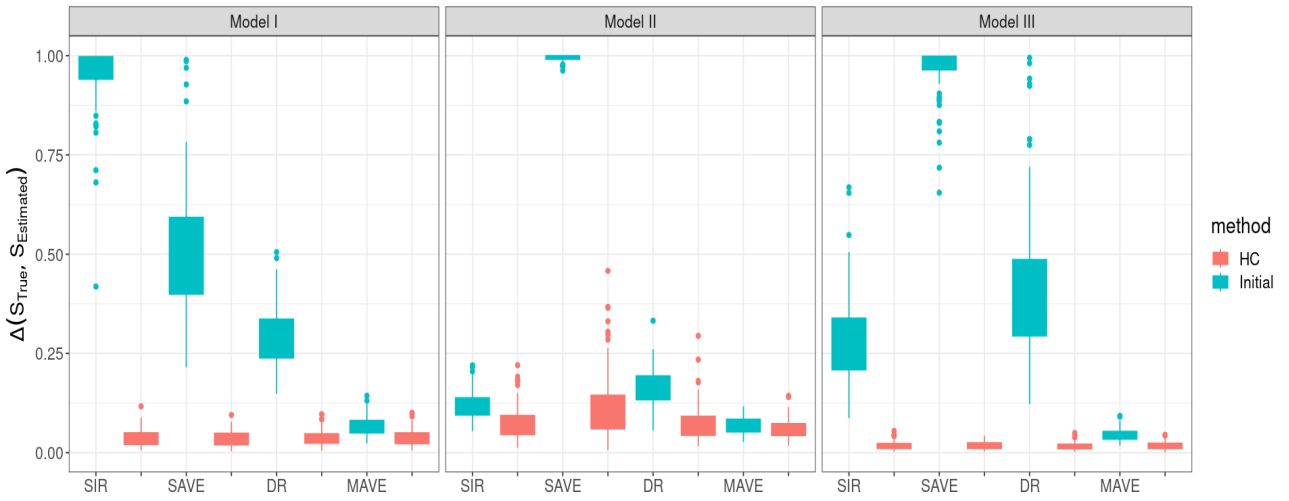


*Figure 1.* Boxplots of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n = 100$ with normal predictors. We compare the performance between SDR Method (aqua blue) and our proposed SDR Method-HC (light coral). As shown, our proposed method consistently outperforms the corresponding SDR methods.

*Table 4.* Real data analysis: Test MSE of house price with SDR methods. The last row is the result from the generalized additive models (GAM) without dimension reduction.

| SDR Methods | MSE: SDR | MSE: SDR-HC |
|---|---|---|
| SIR | 0.313 | **0.243** |
| SAVE | 0.496 | 0.248 |
| DR | 0.314 | 0.264 |
| MAVE | 0.485 | **0.243** |
| DCOV | 0.301 | 0.263 |
| HSIC | 0.257 | 0.252 |
| BCOV | 0.279 | 0.260 |
| GAM(without SDR) | 0.245 | - |

which tells that "house age" is a dominating factor in the single-index nonparametric regression model estimation, $\hat{Y} = \hat{f}(\hat{\eta}^{\top} X)$.

Table 4 shows that our proposed approach can succinctly capture the essential characteristics of predictors in regression using a single-index model, while maintaining regression performance expressed by MSE.

## 6. Discussion

In this work, we introduce a novel approach to recovering the central space by leveraging the Hellinger correlation, specifically designed for scenarios with a structural dimension of one. Our method sets itself apart from existing approaches, such as those dependent on distance covariance and HSIC, by relaxing the stringent requirements of independence assumptions, which frequently present challenges in practical applications. Significantly, our method excels at deriving theoretical results without imposing such technical constraints. Moreover, numerical experiments demonstrate its capability to enhance current existing sufficient dimension reduction methods. Furthermore, the single-index model SDR provides an interpretation of the intrinsic structure of the nonparametric regression model, as demonstrated in the real-world data analysis.

Although the current approach has proven effective, it opens up several important avenues for future work. One promising direction is to extend the application of the proposed method to classification problems. This extension would involve modifying the Hellinger correlation to accommodate categorical variables, such as by leveraging the discrete f-divergence (Geenens, 2020). The question of interest is then to see whether our method can help improve classification accuracy while maintaining theoretical validity under weak assumptions. Another direction for future work is to extend our framework to multi-index models by using a sequential generation of SDR directions as done in Christou (2020) or generalizing the Hellinger correlation to multivariate cases. Additionally, it would be valuable to study the convergence rate of the proposed method, which would provide a deeper understanding of its performance across various settings. We believe that exploring these topics will expand the applicability of our method and make valuable contributions to the field.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## Acknowledgements

We would like to thank four anonymous reviewers and the program chair for the valuable comments and suggestions, which have significantly improved the quality of the paper. For Jun Song, this work is supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. 2022R1C1C1003647, 2022M3J6A1063595, and RS-2023-00219212) and a Korea University Grant (K2402531). Ilmun Kim acknowledges support from the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A4A1033384), the Korean government (MSIT) (RS-2023-00211073), and support from the Yonsei University Research Fund (2022-22-0289).

## References

Adragni, K. P. and Cook, R. D. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.

Bélisle, C. J. Convergence theorems for a class of simulated annealing algorithms on Rd. *Journal of Applied Probability*, 29(4):885–895, 1992.

Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society*, 35: 99–110, 1943.

Cherubini, U., Luciano, E., and Vecchiato, W. *Copula methods in finance*. John Wiley & Sons, 2004.

Chiaromonte, F. and Martinelli, J. Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, 176(1):123–144, 2002.

Christou, E. Central quantile subspace. *Statistics and Computing*, 30(3):677–695, 2020.

Cook, R. D. and Weisberg, S. Sliced Inverse Regression for Dimension Reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

Diaconis, P. and Freedman, D. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984.

Dong, Y. A brief review of linear sufficient dimension reduction through optimization. *Journal of Statistical Planning and Inference*, 211:154–161, 3 2021.

Geenens, G. Copula modeling for discrete random vectors. *Dependence Modeling*, 8(1):417–440, 2020.

Geenens, G. and Lafaye de Micheaux, P. The Hellinger Correlation. *Journal of the American Statistical Association*, 117(538):639–653, 2022.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In *International Conference on Algorithmic Learning Theory*, pp. 63–77, 2005.

Hsueh, H.-M. and Tsai, C.-A. Gene set analysis using sufficient dimension reduction. *BMC Bioinformatics*, 17 (1), February 2016.

Kinney, J. B. and Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9):3354–3359, 3 2014.

Leonenko, N., Pronzato, L., and Savani, V. A class of rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 10 2008.

Li, B. *Sufficient Dimension Reduction*. Chapman and Hall/CRC, April 2018.

Li, B. and Wang, S. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 9 2007.

Li, K.-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414): 316–327, 1991.

Naik, P. A., Hagerty, M. R., and Tsai, C.-L. A new dimension reduction approach for data-rich marketing environments: sliced inverse regression. *Journal of Marketing Research*, 37(1):88–101, 2000.

Nelder, J. A. and Mead, R. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. Ball Covariance: A Generic Measure of Dependence in Banach Space. *Journal of the American Statistical Association*, 115(529):307–317, 1 2020.

Roley, S. S. and Newman, R. M. Predicting Eurasian watermilfoil invasions in Minnesota. *Lake and Reservoir Management*, 24(4):361–369, 2008.

Sheng, W. and Yin, X. Direction estimation in single-index models via distance covariance. *Journal of Multivariate Analysis*, 122:148–161, 11 2013.

Sheng, W. and Yin, X. Sufficient Dimension Reduction via Distance Covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104, 1 2016.

Sklar, M. Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP*, 8(3):229–231, 1959.

Székely, G. J. and Rizzo, M. L. Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236–1265, 12 2009.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 12 2007.

Wang, J. Quantitative trading models based on sufficient dimension reduction and ensemble learning. *Highlights in Business, Economics and Management*, 19:6–16, 2023.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):363–410, 2002.

Xue, Y., Zhang, N., Yin, X., and Zheng, H. Sufficient dimension reduction using Hilbert–Schmidt independence criterion. *Computational Statistics and Data Analysis*, 115:67–78, 11 2017.

Yeh, I.-C. Real estate valuation. UCI Machine Learning Repository, 2018.

Zhang, J. and Chen, X. Robust sufficient dimension reduction via ball covariance. *Computational Statistics and Data Analysis*, 140:144–154, 12 2019.

Zhang, N. and Yin, X. Direction estimation in single-index regressions via Hilbert-Schmidt independence criterion. *Statistica Sinica*, 25(2):743–758, 4 2015.

## A. Additional Simulation Results

The following tables present simulation results when the predictors are non-normal. In most cases, our method significantly enhances existing SDR methods. The specific simulation settings are provided in Section 4.

*Table 5.* Model I: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when predictors are non-normal.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| $n = 100$ | 0.6971 | 0.2674 | 0.6735 | 0.2391 | 0.3958 | 0.2381 | 0.3233 | 0.2609 |
| | (0.2136) | (0.2471) | (0.2009) | (0.2264) | (0.1993) | (0.2279) | (0.3243) | (0.2457) |
| $n = 200$ | 0.7144 | 0.1882 | 0.6396 | 0.1636 | 0.4058 | 0.1804 | 0.1217 | 0.1915 |
| | (0.1396) | (0.1613) | (0.2075) | (0.1299) | (0.1405) | (0.1521) | (0.1517) | (0.1565) |
| $n = 400$ | 0.7284 | 0.1095 | 0.472 | 0.1171 | 0.3929 | 0.1108 | 0.0533 | 0.1193 |
| | (0.1064) | (0.079) | (0.2394) | (0.0925) | (0.1284) | (0.0851) | (0.0264) | (0.0977) |

Table 5 shows the experimental results under Model I with the non-normal predictors. It presents $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ when we use SIR (Li, 1991), SAVE (Cook & Weisberg, 1991), DR (Li & Wang, 2007), and MAVE (Xia et al., 2002). As in the main text, the SIR-HC, SAVE-HC, DR-HC, MAVE-HC refer to our approaches, using SIR, SAVE, DR, and MAVE, respectively, as their initial values in the iterations. One can observe that SIR fails to recover the central subspace as previously mentioned for cases with normal predictors. In contrast, our proposed method based on SIR successfully discloses the central subspace.

*Table 6.* Model II: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when predictors are non-normal.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| $n = 100$ | 0.2287 | 0.0666 | 0.9200 | 0.1190 | 0.4555 | 0.1049 | 0.0686 | 0.0377 |
| | (0.1429) | (0.0679) | (0.0583) | (0.144) | (0.2447) | (0.1307) | (0.0327) | (0.021) |
| $n = 200$ | 0.1053 | 0.0193 | 0.9067 | 0.0557 | 0.3717 | 0.0425 | 0.0296 | 0.0121 |
| | (0.0476) | (0.0214) | (0.1324) | (0.0965) | (0.2233) | (0.0712) | (0.0125) | (0.0091) |
| $n = 400$ | 0.0881 | 0.0069 | 0.5365 | 0.0162 | 0.359 | 0.0126 | 0.0148 | 0.004 |
| | (0.0524) | (0.0074) | (0.3664) | (0.0273) | (0.2045) | (0.0243) | (0.0063) | (0.0027) |

*Table 7.* Model III: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when predictors are non-normal.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| $n = 100$ | 0.2457 | 0.0421 | 0.7842 | 0.0393 | 0.4237 | 0.0379 | 0.0286 | 0.0363 |
| | (0.1431) | (0.0286) | (0.2328) | (0.0236) | (0.2585) | (0.0272) | (0.0178) | (0.0251) |
| $n = 200$ | 0.2044 | 0.0298 | 0.8084 | 0.0292 | 0.3501 | 0.0317 | 0.0178 | 0.0288 |
| | (0.0997) | (0.0235) | (0.1985) | (0.0226) | (0.2204) | (0.0255) | (0.0097) | (0.0225) |
| $n = 400$ | 0.1889 | 0.0229 | 0.7952 | 0.0237 | 0.3232 | 0.0246 | 0.0117 | 0.0226 |
| | (0.0824) | (0.0186) | (0.1989) | (0.0176) | (0.1856) | (0.0175) | (0.0072) | (0.0175) |

Table 6 and 7 show the experimental results under Model II and Model III with the non-normal predictors. Similar to normal cases, our method can effectively capture the central space and significantly improves its accuracy.

Tables 8–10 show the experimental results using modern SDR methods as the initial values. We compared three existing methods (via distance covariance (Sheng & Yin, 2013), via HSIC (Zhang & Yin, 2015) and via Ball covariance (Zhang & Chen, 2019)) that capture the central subspace by maximizing dependency measures. The simulation settings are same as those in Tables 1–3. The predictors follow normal distributions. The true $\eta$s and models are detailed in Section 4. Our proposed method enhances outcomes by using initial values provided by these three methods. In most cases, the standard deviation has also decreased. One can verify that our method still improves the results, even though the results of the existing method were already promising.

*Table 8.* Model I: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ where modern methods are initial methods.

| | DCOV | DCOV-HC | HSIC | HSIC-HC | BCOV | BCOV-HC |
|---|---|---|---|---|---|---|
| n = 100 | 0.1521 | 0.0373 | 0.1578 | 0.0379 | 0.1425 | 0.0339 |
| | (0.1744) | (0.0224) | (0.1694) | (0.0215) | (0.1943) | (0.0191) |
| n = 200 | 0.0966 | 0.0232 | 0.0948 | 0.0224 | 0.0747 | 0.0227 |
| | (0.159) | (0.0131) | (0.1293) | (0.0141) | (0.1628) | (0.0121) |
| n = 400 | 0.0732 | 0.0177 | 0.0804 | 0.0182 | 0.0556 | 0.0184 |
| | (0.1617) | (0.0086) | (0.1608) | (0.0095) | (0.1654) | (0.0089) |

*Table 9.* Model II: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ where modern methods are initial methods.

| | DCOV | DCOV-HC | HSIC | HSIC-HC | BCOV | BCOV-HC |
|---|---|---|---|---|---|---|
| n = 100 | 0.1632 | 0.0843 | 0.1077 | 0.0680 | 0.1087 | 0.0660 |
| | (0.076) | (0.0539) | (0.0332) | (0.0354) | (0.0422) | (0.0365) |
| n = 200 | 0.0997 | 0.0411 | 0.0743 | 0.0373 | 0.0396 | 0.0307 |
| | (0.0291) | (0.021) | (0.0219) | (0.0173) | (0.0153) | (0.0155) |
| n = 400 | 0.0657 | 0.0219 | 0.0469 | 0.0204 | 0.0289 | 0.0163 |
| | (0.0549) | (0.011) | (0.0131) | (0.0089) | (0.0983) | (0.007) |

*Table 10.* Model III: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ where modern methods are initial methods.

| | DCOV | DCOV-HC | HSIC | HSIC-HC | BCOV | BCOV-HC |
|---|---|---|---|---|---|---|
| n = 100 | 0.1459 | 0.0194 | 0.1223 | 0.0177 | 0.0594 | 0.0201 |
| | (0.0452) | (0.0103) | (0.0399) | (0.0104) | (0.0284) | (0.0116) |
| n = 200 | 0.0873 | 0.0126 | 0.075 | 0.0114 | 0.0204 | 0.0136 |
| | (0.0252) | (0.0067) | (0.0201) | (0.0066) | (0.007) | (0.006) |
| n = 400 | 0.0594 | 0.0087 | 0.0503 | 0.0090 | 0.0114 | 0.009 |
| | (0.0143) | (0.0039) | (0.0127) | (0.0043) | (0.0035) | (0.0041) |

Tables 11–13 show the experimental results when $\eta$ is not sparse. The sparsity of the direction vector does not affect our proposed method as well as the other existing methods. We changed only the true $\eta$s as follows:

$$\text{Model I} : \eta = (1,1,1,1,1,1,1,1,1,1)^{\mathsf{T}}/\sqrt{10}.$$
$$\text{Model II} : \eta = (1,1,1,-1,-1,-1,-1,1,1,-1)^{\mathsf{T}}/\sqrt{10}.$$
$$\text{Model III} : \eta = (3,-1,4,-2,-4,5,1,-3,-5,2)^{\mathsf{T}}/\sqrt{110}.$$

All other conditions, including the models and the distribution of predictors, remain the same as described in Section 4.

*Table 11.* Model I: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when true $\eta$ is non-sparse.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| n = 100 | 0.9462 | 0.8865 | 0.5093 | 0.6764 | 0.2946 | 0.2274 | 0.0534 | 0.0328 |
| | (0.0903) | (0.1755) | (0.185) | (0.3597) | (0.0766) | (0.2351) | (0.015) | (0.0112) |
| n = 200 | 0.8857 | 0.8435 | 0.3138 | 0.1749 | 0.1821 | 0.0678 | 0.0246 | 0.0152 |
| | (0.154) | (0.1636) | (0.1186) | (0.2072) | (0.0442) | (0.0471) | (0.0061) | (0.0046) |
| n = 400 | 0.8804 | 0.6878 | 0.1819 | 0.0564 | 0.1249 | 0.0403 | 0.013 | 0.0085 |
| | (0.1703) | (0.3057) | (0.0538) | (0.0415) | (0.0349) | (0.0325) | (0.0032) | (0.0026) |

*Table 12.* Model II: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when true $\eta$ is non-sparse.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| n = 100 | 0.1554 | 0.1223 | 0.9922 | 0.7911 | 0.5444 | 0.5769 | 0.0725 | 0.0741 |
| | (0.0431) | (0.0496) | (0.0095) | (0.1455) | (0.1962) | (0.3026) | (0.0259) | (0.0358) |
| n = 200 | 0.0992 | 0.0677 | 0.8597 | 0.6362 | 0.4024 | 0.2929 | 0.0427 | 0.0372 |
| | (0.0272) | (0.0307) | (0.2312) | (0.255) | (0.1727) | (0.2632) | (0.0137) | (0.0194) |
| n = 400 | 0.0519 | 0.0623 | 0.0696 | 0.0679 | 0.0704 | 0.0691 | 0.0378 | 0.0594 |
| | (0.0109) | (0.0189) | (0.0181) | (0.0228) | (0.0168) | (0.0194) | (0.0091) | (0.0179) |

*Table 13.* Model III: Mean and standard deviation (with parentheses) of $\Delta(\mathcal{S}_{True}, \mathcal{S}_{Estimated})$ over 100 samples of size $n$ when true $\eta$ is non-sparse.

| | SIR | SIR-HC | SAVE | SAVE-HC | DR | DR-HC | MAVE | MAVE-HC |
|---|---|---|---|---|---|---|---|---|
| n = 100 | 0.1405 | 0.0871 | 0.9903 | 0.763 | 0.194 | 0.1012 | 0.0371 | 0.0398 |
| | (0.0568) | (0.0497) | (0.0133) | (0.1594) | (0.0831) | (0.098) | (0.0131) | (0.0111) |
| n = 200 | 0.0853 | 0.0409 | 0.8053 | 0.5614 | 0.1198 | 0.0542 | 0.0176 | 0.0243 |
| | (0.0289) | (0.018) | (0.2765) | (0.256) | (0.0417) | (0.0273) | (0.0047) | (0.0073) |
| n = 400 | 0.0624 | 0.03 | 0.0893 | 0.0373 | 0.0808 | 0.0338 | 0.0109 | 0.0174 |
| | (0.0201) | (0.0116) | (0.0334) | (0.0201) | (0.0244) | (0.0168) | (0.0029) | (0.0057) |

The overall behavior of the experimental results with non-sparse direction vectors is not significantly different from those in Tables 1–3. We observe that, as the sample size increases, our proposed method detects the true directions with high accuracy.