entropy

MDPI

*Article*

# Sufficient Dimension Reduction: An Information-Theoretic Viewpoint

**Debashis Ghosh**

Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO 80045, USA; debashis.ghosh@cuanschutz.edu

**Abstract:** There has been a lot of interest in sufficient dimension reduction (SDR) methodologies, as well as nonlinear extensions in the statistics literature. The SDR methodology has previously been motivated by several considerations: (a) finding data-driven subspaces that capture the essential facets of regression relationships; (b) analyzing data in a 'model-free' manner. In this article, we develop an approach to interpreting SDR techniques using information theory. Such a framework leads to a more assumption-lean understanding of what SDR methods do and also allows for some connections to results in the information theory literature.

**Keywords:** central subspace; information bottleneck; single-index model

## 1. Introduction

In statistical modeling, a key challenge is to determine appropriate transformations of the data that can reduce its dimension while at the same time capturing the essential information in the regression relationship between a set of covariates and a response variable. To this end, there has been a field of statistics, termed sufficient dimension reduction (SDR), that has sought to develop a methodology with this goal in mind. Broadly speaking, sufficient dimension reduction represents a class of 'model-free' methodologies that seek to find directions in the data that can capture the essential information in the regression relationship previously mentioned. An excellent recent monograph on the topic can be found in [1].

Historically, the basis for sufficient dimension reduction methods was the observation by authors, such as Brillinger [2] and Li and Duan [3], that regression parameters estimated by ordinary least squares were consistent, up to a constant, for their population counterparts in a generalized single-index model. This result required an assumption on the covariates being elliptically symmetric, which has been reframed into the current sufficient dimension reduction literature as the so-called linearity assumption. More recent formulations for sufficient dimension reduction have postulated the existence of a central subspace; subsequently, the goal of sufficient dimension reduction methods is to estimate the basis vectors of the central subspace. There now exists a wide variety of techniques available for estimation in sufficient dimension reduction; we will provide a review of such methods in Section 2.1.

In this article, we propose a new interpretation for sufficient dimension reduction based on conditional independence assumptions. Using graphical models, we are able to connect sufficient dimension reduction methods to information bottleneck theory [4]. The information bottleneck methodology was pioneered by the late Naftali Tishby and seeks to develop a 'short code for X that preserves the maximum information about Y' [4]. Information bottleneck typically formulates an optimization problem that seeks to find a compressed representation that minimizes information loss while imposing a penalty related to the expected distortion of the compression. The compression is the 'bottleneck' in the term 'information bottleneck'. The optimization is solved using calculus of variations and leads to a set of self-consistent equations for finding optimal codes that are related to proposals by Blahut [5] and Arimoto [6]. The information bottleneck approach has been

check for updates

**Citation:** Ghosh, D. Sufficient Dimension Reduction: An Information-Theoretic Viewpoint. *Entropy* **2022**, *24*, 167. https://doi.org/10.3390/e24020167

Academic Editor: Sangun Park

Received: 3 December 2021
Accepted: 18 January 2022
Published: 22 January 2022

applied to a variety of problems in machine learning, such as document clustering [7,8], multivariate density estimation [9] and deep learning [10,11].

The interpretation developed in this paper allows us to demonstrate the following:

(a). We can view sufficient dimension reduction as a means of preserving information that is relevant to a response variable. It can be interpreted as performing the information bottleneck in two directions.

(b). Conversely, we will see that the information bottleneck is performing sufficient dimension reduction in a certain sense.

(c). By moving to mutual information, we can relax some of the distributional assumptions needed for sufficient dimension reduction in a manner different from that in [12–16]. This direction is a departure from the viewpoint that SDR serves as a means to estimate a target parameter, typically the span of basis vectors of the central subspace.

(d). In the case of Gaussian variables, we can develop a method for identifying ‘phase transitions’ in the structural dimension of central subspaces by expanding the work of [17] to handle sufficient dimension reduction procedures.

While many of the information-theoretic results are well-known to the information theory community, their embedding and merging with the literature on sufficient dimension reduction will be novel to statisticians.

The closest statistical work in nature to ours is that of Wang et al. [18]. They leverage the Hellinger integral of order two [19], which is related to the Kullback–Leibler divergence, an important quantity in information theory. Wang et al. [18] define subspace-based information measures using the Hellinger integral and demonstrate that a central subspace preserves information on this scale. For its estimation, they use a nonparametric regression approach that bears a resemblance to the minimum average variance estimation approach of Xia et al. [12]. The idea of using the Kullback–Leibler divergence for the optimization and estimation of the central subspace and other measures of association was used by Yin and co-authors in a series of papers [20–23]. We will also note work by Cook and Ni [24], who use a minimum discrepancy approach for finding the central subspace, and the work of Yao et al. [25], who developed a sufficient reduction procedure using the Fisher information metric, which can also be shown to be connected to Kullback–Leibler divergence. Finally, a device we use in the paper is graphical models, and we note recent work by [26].

The outline of this paper is as follows. We review the literature on sufficient dimension reduction, as well as pointing out some limitations, in Section 2. Section 3 seeks to develop connections between sufficient dimension reduction and information theory using graphical models. We focus on the Gaussian information bottleneck [17] and its relationship to SDR in Section 4. We illustrate the methodology with application to a dataset in Section 5. Section 6 concludes with some discussion.

## 2. Background and Preliminaries

### 2.1. Data Structures and Review of Dimension Reduction Methods

Much of the material presented here is expounded upon in the monograph by Li [1]. Let the data be represented as $(Y_i, Z_i)$, $i = 1, \ldots, n$, a random sample from the joint distribution $(Y, Z)$, where $Y$ denotes the response of interest and $Z$ is a $p$-dimensional vector of covariates. Suppose we formulate the following regression model for $Y$ given $Z$:

$$E(Y \mid Z) = g(\beta_1' Z, \beta_2' Z, \ldots, \beta_k' Z, u), \tag{1}$$

where $\beta_j$ $(j = 1, \ldots, k)$ are $p$-dimensional vectors of unknown regression coefficients, $u$ is an error term, and $g$ is an unspecified monotonic link function. Because of the presence of the parametric components involving $\beta_j$, as well as the nonparametric specification of the link function, model (1) is semiparametric. Note that when $k = 1$, model (1) reduces to a single-index model [27]. In addition, model (1) can accommodate non-homoskedasticity in the error term if the variance depends on $\beta_j' Z$.

The starting point of dimension reduction methods is the conditional independence of $Y$ and $Z$ given $E(Y \mid Z)$. We define two random variables, $A$ and $B$, to be conditionally independent given $C$ if

$$P(A|B,C) = P(A|C).$$

We will use the notation $A \perp\!\!\!\perp B|C$ to represent conditional independence. An implication of model (1) being true is that there exists a $p \times k$ matrix $\mathbf{B}$, where

$$Y \perp\!\!\!\perp Z|\mathbf{B}'Z. \tag{2}$$

Another way of stating (2) is that the projection $\mathbf{B}'Z$ provides a sufficient data reduction and contains the essential information about the relationship between $Z$ and $Y$. More generally, we can define a projection operator $P_{\mathbf{B}}$ to be the symmetric and idempotent operator onto the subspace spanned by the columns of $\mathbf{B}$. Then, (2) can be re-expressed as

$$Y \perp\!\!\!\perp Z \mid P_B Z. \tag{3}$$

If (3) holds, then it also holds for any subspace of $\tilde{B}$ such that the span of $\mathbf{B}$ is the same as the span of $\tilde{C}$. Let $S(\mathbf{B})$ be the subspace generated by the columns of $\mathbf{B}$. Let $S_{Y|Z}$ denote the intersection of all possible subspaces; if $S_{Y|Z}$ also satisfies (3), then we will refer to $S_{Y|Z}$ as the central subspace [28]. We will assume throughout that the central subspace exists [28–30]. In the classical presentation for sufficient dimension reduction methodology, the parameter has been defined to be the span of $S_{Y|Z}$. In other words, if $\mathbf{v}_1, \dots, \mathbf{v}_K$ denote the basis vectors for $S_{Y|Z}$, then

$$S_{Y|Z} \equiv \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_K)$$

is the target of sufficient dimension reduction procedures. Thus, there is an estimand that is often targeted by sufficient dimension reduction procedures.

We assume, without a loss of generality, that $Z$ has a mean zero vector and covariance matrix equal to the identity matrix. One key assumption necessary for the implementation of one class of sufficient dimension reduction procedures is that the distribution of $Z$, conditional on $P_{\mathbf{B}}Z$, satisfies a conditional linearity in the mean, i.e.,

$$E(Z \mid P_{\mathbf{B}}Z) = P_{\mathbf{B}}Z. \tag{4}$$

Assumption (4) pertains to the marginal distribution of $Z$ and means that all the information about $Z$ is contained in its projection onto the subspace spanned by $\mathbf{B}$. One class of distributions that satisfies the linearity condition is the family of elliptically symmetric distributions. This includes distributions, such as the multivariate normal distributions and scale mixtures of multivariate normal distributions.

As mentioned in the Introduction, there are many algorithms available for estimating the basis vectors of the central subspace. We describe the implementation of sliced inverse regression proposed by Li [28].

(a). 'Slice' the response variable $Y$ into $J$ slices, denoted as $\mathcal{Y}_1, \dots, \mathcal{Y}_J$;

(b). Standardize the predictor observations as

$$\tilde{Z}_i = \hat{\Sigma}^{-1/2}(Z_i - \hat{\mu}), (i = 1, \dots, n),$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance matrices of $Z_1, \dots, Z_n$;

(c). Calculate sample mean estimates within slices: $\bar{Z}_j = n_j{}^{-1} \sum_{i=1}^{n} I(Y_i \in \mathcal{Y}_j)\tilde{Z}_i$, where $n_j = \sum_{i=1}^{n} I(Y_i \in \mathcal{Y}_j), j = 1, \dots, J$;

(d). Estimate the population covariance matrix of $Z$ given $Y$ by

$$\hat{\Theta} = \sum_{j=1}^{J} \frac{n_j}{n} \bar{Z}_j \bar{Z}_j';$$

(e). Compute the eigenvalues of $\hat{\Theta}$. These are the estimates of the basis vectors for the central subspace.

This algorithm is termed 'inverse regression' because effectively, information on the 'backwards regression' $E(Z \mid Y)$ is being estimated here rather than the 'forward regression' $E(Y \mid Z)$. Li [28] argues that this approach circumvents the usual issue of the curse of dimensionality. Other advantages of the sliced inverse regression algorithm are that it avoids multivariate nonparametric smoothing and is quite easy to fit.

The validity of the sliced inverse regression algorithm for estimating the central subspace relies on the linearity assumption. There has been much work on developing alternative estimation procedures that seek to relax the linearity assumption. For example, Xia et al. [12] propose the minimum average variance estimation procedure, which relies on a combination of nonparametric smoothing with weighted least squares. Since it involves nonparametric regression, its convergence depends on an appropriate rate of convergence for the bandwidth in conjunction with the sample size converging to infinity. Cook and Ni [24] proposed a minimum discrepancy method in which sufficient dimension reduction is characterized using an objective function approach. This leads to an alternating least squares algorithm for the estimation of the central subspace.

Many of the sufficient dimension-reduction methods can be viewed as solving the following eigenvalue/eigenvector problem:

$$\mathbf{A}\mathbf{b}_j = \lambda_j \Sigma_Z \mathbf{b}_j, \tag{5}$$

for $j = 1, \ldots, k$, where $\Sigma_Z$ denotes the covariance matrix of $Z$, and $(\lambda_j, b_j)$ denotes the eigenvalue/eigenvector pairs. We say that the matrix pair $(\mathbf{A}, \Sigma_Z)$ is a generalized eigenvalue solution (GES) if it satisfies (5). This is discussed at length in the monograph by Li [1]. Note that typically, the solutions to (5) are returned as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k,$$

and $\mathbf{b}_j' \Sigma_Z \mathbf{b}_j = 1$ for $j = 1, \ldots, k$. The choice of $\mathbf{A}$ in (5) depends on the particular sufficient dimension reduction algorithm that is used. For example, in sliced inverse regression, $\mathbf{A}$ would represent the covariance matrix of the slice means. For principal Hessians directions [31], $\mathbf{A}$ in (5) is taken to be a weighted covariance matrix of the response to $Z$.

The matrix formulation in (5) allows for immediate generalizations to nonlinear versions of sufficient dimension. This can be done by replacing $\mathbf{A}$ in (5) with a so-called 'kernelized' matrix computed using inner products of covariates mapped to higher-dimensional spaces. Such methods are related to the procedures in Wu et al. [32], Fukumizu et al. [14] and Lee et al. [16].

### *2.2. Limitations of Sufficient Dimension Reduction*

As mentioned above, one of the key assumptions in applying sufficient dimension reduction methodology is termed the linearity condition. A sufficient condition for this to hold is that the predictor variables of interest follow an elliptically contoured distribution. Distributions that satisfy elliptical symmetry include the multivariate normal distribution and the multivariate $t$-distribution. One of the main criticisms leveled against the sufficient dimension reduction methods is that this assumption will not be satisfied in practice. For example, if covariates are discrete, then this will violate the linearity condition. Many authors invoke the theoretical results of Hall and Li [33], which suggests that in an asymptotic framework, the linearity condition will hold. An alternative approach has been to develop generalizations of the sufficient dimension reduction methodology that relax the linearity condition. Such approaches can be found in proposals, such as Chiaromonte et al. [34], Fukumizu et al. [13,14], Li et al. [15] and Lee et al. [16].

The other issue with sufficient dimension reduction methods involves the identification of the basis of $S_{Y|Z}$, which is referred to as the directions of the central subspace. These vectors are not estimable in the situation where the components of $Z$ are discrete.

Such variables arise routinely in biomedical, sociological and demographic studies (e.g., race/gender), and this limitation makes the use of sufficient dimension reduction methods challenging. In an important work, Chiaromonte et al. [34] developed an approach to sufficient dimension reduction with categorical predictors. The idea is to perform the sliced averaging of the continuous covariates within each of the levels defined by the combination of the categorical variables. Then, the level-specific covariance matrices are pooled, and the directions are estimated using spectral decomposition, similar to the description of sliced inverse regression in Section 2.1.

## 3. Graphical Models, Connections and Information Theoretic Results

To link sufficient dimension reduction methods to the information bottleneck, we will now introduce some concepts from graph theory and graphical models [26,35]. A graph $G = (V, E)$ consists of a set of vertices $V$ and a collection of edges $E$. Here, $V \equiv \{v_1, \ldots, v_m\}$ denotes the collection of $m$ vertices and the edges $E$ consist of two-element subsets of the power set of $V$ that denote edges between vertices. To simplify the discussion, we will assume that there are no edges from a vertex to itself, i.e., no self-loops. Graphs whose edges have more than two elements are referred to as hypergraphs [36] and will not be considered further in the paper. The vertices represent random variables, and the edges are used to specify dependencies between the random variables. There are two types of edges we will consider here between vertices $v_1$ and $v_2$. A directed edge is denoted by $v_1 \to v_2$ and implies that $v_1$ affects $v_2$ and not vice versa. An undirected edge is denoted by $v_1 - v_2$ and is equivalent to $v_1 \to v_2$ and $v_2 \to v_1$. Thus for undirected edges, $v_1$ and $v_2$ simultaneously affect each other. We define the parents of a vertex $v$ by

$$\mathrm{pa}(v) = \{u \in V : \exists \text{ a directed edge from } u \text{ to } v\}.$$

It is a well-known fact that for an acyclic directed graph [26,35], one can factorize the joint distribution of random variables defined on the graph $G$ as

$$f(X_u : u \in G) = \prod_{v \in V} p(x_v | x_s,\ s \in pa(v)). \tag{6}$$

The final graphical model concept we will need is that of d-separation [35]. If G is a directed graph in which $X$, $Y$ and $Z$. are a disjoint sets of vertices, then $X$ and $Y$ are d-separated by $Z$ in $G$ if and only if every path from a vertex in $X$ to a vertex in $Y$ is intercepted by a vertex in $Z$.

We can see that assumption (3) corresponds to the following graphs

$$\begin{array}{lll} Z & \to P_B Z & \to Y \\ Z & \leftarrow P_B Z & \leftarrow Y \\ Z & \to P_B Z & \leftarrow Y \\ Z & \leftarrow P_B Z & \to Y \end{array} \tag{7}$$

This follows from using the definition of undirected graphs and conditional independence. Similarly, the information bottleneck approach works with the graph

$$Z \to T \to Y \tag{8}$$

The comparison of (7) and (8) offers the following insights. First, the central subspace performs d-separation of $Z$ and $Y$. Similarly, the role of $T$ in the information bottleneck framework is to intercept paths between $Z$ and $Y$. This leads us to the following result, which will be new to statisticians:

**Proposition 1.** *The central subspace can also serve as an information bottleneck.*

**Proof.** The proof of the proposition follows by observing that the graph in (8) is a subgraph of the graphs in (7). □

**Remark 1.** *Returning to the work of Wang et al. [18], the graphical representation in (7) makes sufficient dimension reduction integrating the forward and backward regressions. The graphs in (7) are precisely the forward and backward regression that Wang and colleagues speak of. They can also be viewed as 'forward' and 'backward' information bottlenecks. Thus, we see that sufficient dimension reduction is attempting to simultaneously perform a forward and reverse information bottleneck, while information bottlenecks themselves operate in the forward direction.*

*Based on the proposition, we observe that the role of the central subspace in sufficient dimension reduction plays a role akin to the information bottleneck. Using the viewpoint of information theory, we can interpret the goal of sufficient dimension reduction as one of information compression. This allows the use of these methods even in situations when the central subspace will not be estimable.*

To make the idea concrete, we will be interpreting $P_B Z$ as a random variable in the rest of the section. We will further assume that $Z$ and $Y$ are discrete random variables that are potentially multivariate. The entropy of $Z$ is defined by

$$H(Z) = -\log \sum_{z \in \mathcal{Z}} p(z) \log p(z), \tag{9}$$

where $\mathcal{Z}$ is the range of $Z$ and $p(z)$ denotes the probability mass function. Similarly, we can define the mutual information of $Z$ and $Y$ as

$$I(Z;Y) = \sum_{z,y} p(z,y) \log \frac{p(z,y)}{p(z)p(y)} \tag{10}$$

which extends upon (9) in a natural way. Mutual information measures the dependence between two random variables. It has the following properties: (a) it is symmetric in $Z$ and $Y$; (b) it is nonnegative; (c) it is equal to zero if and only if $Z$ and $Y$ are independent.

A comprehensive overview of entropy and mutual information can be found in Cover and Thomas [37]. To keep the discussion self-contained, we now provide a summary of many basic properties of entropy and mutual information. Further details can be found in Chapter 2 of Cover and Thomas [37].

**Property 1.**

1. $I(Z;Y) = H(Z) - H(Z|Y)$, *where* $H(Z|Y) = -\sum_{z,y} p(z,y) \log p(z|y)$.
2. $I(Z;Y) = H(Y) - H(Y|Z)$.
3. $I(Z;Y) = I(Y;Z)$.
4. $I(Z;Y) = H(Z) + H(Y) - H(X,Y)$.
5. $H(X,Y) = H(X) + H(Y|X)$.

We note that the last property is typically referred to as the chain rule for entropy and can be extended to more than two random variables.

For the graphs considered in (7) and (8), we need to consider conditional versions of mutual information. This is given to us by Equation (2.61) of Cover and Thomas [37].

**Definition 1.** *The conditional mutual information of $Z$ and $Y$ given $W$ is defined as*

$$I(Z;Y|W) = \sum_{z,y,w} p(z,y,w) \log \frac{p(z,y|w)}{p(z|w)p(y|w)}.$$

*Finally, we will need one more definition from Chapter 2.8 of Cover and Thomas [37].*

**Definition 2** (p. 34 of [37])**.** *$Z, Y$ and $W$ form a Markov Chain, denoted as $Z \to W \to Y$ if the conditional distribution of $Y$ given $W$ and $Z$ only depends on $Z$.*

We assume a reversible Markov Chain so that $Z \to W \to Y$ and $Y \to W \to Z$ are treated as equivalent. Thus, the reversibility of the Markov Chain allows us to conceptually drop the directionality in DAGs, which becomes in line with the conditional independence assumptions outlined in Section 2.1. We have the following celebrated result from information theory, the data-processing inequality (p. 34 of [37]):

**Theorem 1.** *If $Z \to W \to Y$, then $I(W;Z) \geq I(Y;Z)$.*

The data processing inequality guarantees equality if and only if $Z$ and $Y$ are independent given $W$. We can now take these results and apply them to the graphs for sufficient dimension reduction.

**Corollary 1.** *Assumption (3) is equivalent to $I(Z;P_BZ) = I(Z;Y)$. This corollary also relates to Theorem 1 of Wang et al. [18]. This formalizes the proposition earlier in the paper.*

**Remark 2.** *Note that we have rephrased the central subspace as a random variable that attempts to minimize an information-based criterion. Thus, we get away from the traditional viewpoint where we view the goal of sufficient dimension reduction as targeting the span of the central subspace. Doing so provides another justification for the use of sufficient dimension reduction. This is in the spirit of 'assumption-lean inference' [38] in which the goal is to have available statistical methods that can be useful even when a true model or parameter does not exist.*

**Remark 3.** *The mutual information is intimately related to the Kullback–Leibler divergence of two probability measures. A nice overview of how Kullback–Leibler divergences are related to information theoretic quantities can be found in [19]. We will explore the link between sufficient dimension reduction methods and Kullback–Leibler divergences in future work.*

## 4. The Case of Gaussian Variables

In most problems involving the information bottleneck, one can use the Blahut–Arimoto algorithm [5,6], which is an iterative algorithm that involves repeated projection operations. In this section, we study a noniterative information bottleneck algorithm by Chechik et al. [17]. They deal with the situation of $Z$ and $Y$ having a joint Gaussian distribution and show that one can use an eigenvector/eigenvalue decomposition of certain matrices to achieve an information bottleneck. We then show how to relate this to several sufficient dimension reduction procedures.

Chechik et al. [17] considered the situation of $(Z,Y)$ having a joint Gaussian or multivariate normal distribution. Without loss of generality, we will assume a mean of zero throughout the section. The goal of the Gaussian information bottleneck is to find a mapping from $Z$ to $T$, such that the information content of $Z$ is sufficiently compressed while at the same time maintaining its association with $Y$. Formally, the Gaussian information bottleneck involves the minimization of

$$\mathcal{L} \equiv I(Z;T) - \beta I(T;Y)$$

over matrices $\mathbf{A}$ and $\Sigma_e$, where

$$T = \mathbf{A}Z + e, \quad e \sim MVN(0, \Sigma_e), \tag{11}$$

and $MVN(0,\Sigma)$ denotes a multivariate normal distribution with mean zero vector and covariance matrix $\Sigma$. Note that in (11), we assume that $e$ is independent of $Z$. Because of the linearity of $T$ in $Z$ in (11), $T$ will have a multivariate Gaussian distribution with mean zero vector and covariance matrix $\mathbf{A}\Sigma_Z\mathbf{A}' + \Sigma_e$. Chechik et al. [17] prove the following theorem.

**Theorem 2** (Theorem 3.1. of [17])**.** *The optimal solution to the Gausssian information bottleneck problem (11) for a given $\beta$ is given by $\Sigma_e^{opt} = \mathbf{I}$ and*

$$A = \begin{cases} [\mathbf{0}' \cdots \mathbf{0}'] \text{ if } 0 \le \beta \le \beta_1^* \\ [\gamma_1 \mathbf{v}_1' \;\; \mathbf{0}' \cdots] \text{ if } \beta_1^* < \beta \le \beta_2^* \\ [\gamma_1 \mathbf{v}_1' \;\; \gamma_2 \mathbf{v}_2' \;\; \mathbf{0}' \cdots] \text{ if } \beta_2^* < \beta \le \beta_3^* \\ \vdots \end{cases}$$

*where $\gamma_1, \ldots, \gamma_n$ are functions of the eigenvalues $\alpha_1, \ldots, \alpha_n$ of $\Sigma_{Z|Y}\Sigma_Z^{-1}$, defined as*

$$\gamma_i = \sqrt{\frac{\beta(1-\alpha_i)-1}{\alpha_i r_i}},$$

*$r_i = \mathbf{v}_i'\Sigma_Z \mathbf{v}_i$ and*

$$\beta_i^* = (1-\alpha_i)^{-1}, i = 1, \ldots, n.$$

*Thus, the theorem demonstrates the tradeoff between compression and its associated cost. If $\beta$ is smaller than $\beta_1^*$, then the cost is too high, and the optimal solution is the zero matrix. Otherwise, we see that we can start to identify subspaces for larger values of $\beta$ associated with the eigenvectors of $\Sigma_{Z|Y}\Sigma_Z^{-1}$. We also see a transition in terms of the dimensions of the subspaces spanned by $\mathbf{v}_i$ as $\beta$ increases. There are also discrete jump points for $\beta$. We refer to the result of Theorem 2 as the Gaussian Information Bottleneck Theorem.*

Note that the theorem also involves solving the eigenvalue/eigenvector decomposition via the equation

$$\Sigma_{Z|Y}\mathbf{v} = \lambda \Sigma_Z \mathbf{v}.$$

Comparing the structure of this equation to (5), we see that $(\Sigma_{Z|Y}, \Sigma_Z)$ can be viewed as a GES. The difference between the typical GES solution with the result of Theorem 2 is the order in which eigenvalues appear. For GES, they occur in descending order, while for Theorem 2, they are in ascending order. Using the link of Theorem 2 to generalized eigenvector solutions, we have the following result for sufficient dimension reduction methodology [17] to demonstrate the following result.

**Proposition 2.** *The result of the Gaussian information bottleneck theorem holds*

*(a).* *for sliced inverse regression with $Cov\{E(Z|Y), \Sigma_Z\}$ as a GES;*

*(b).* *for partial inverse regression with $Cov\{\Sigma_{ZF}\Sigma_{FF}^{-1}\Sigma_{FZ}, \Sigma_Z\}$ as a GES, where for a known transformation of $Y$, $F(Y)$,*

$$\begin{aligned} \Sigma_{FF} &= VarF(Y) \\ \Sigma_{FZ} &= Cov\{F(Y), Z\} \end{aligned}$$

*(c).* *for sliced average variance estimation [12] with $Cov\{\Sigma_Z - Var(Z|Y), \Sigma_Z\}$ as a GES;*

*(d).* *for principal Hessians directions [31] with $Cov\{\Sigma_{ZZY}, \Sigma_Z\}$ as a GES, where*

$$\Sigma_{ZZY} = E(ZZ'Y)$$

**Proof.** All of these results follow by defining the GES equivalences as found in Li [1]. □

The proposition affords us new insights into how to view the information compression/basis calculation for several existing sufficient dimension reduction procedures from the information bottleneck viewpoint.

We note that if we sort the eigenvectors in descending order, the problem of selecting which index to stop is precisely that of selecting the dimension of the central subspace. This is an important problem for which there have been several approaches in the literature. Ye and Weiss [39] proposed an approach to the selection using the nonparametric bootstrap. Recently, Luo and Li [40] proposed the ladle approach, which used the bootstrap but combined information from both the eigenvalues and the eigenvectors of the central subspace to determine the dimension of the central subspace. Another recent innovation by Luo and Li [41] was to augment the predictor matrix with noise variables, which is in the spirit of the recent, popular 'knockoffs' approach in statistics [42]. One sees that the problem of order determination of the central subspace is dual to the Gaussian information bottleneck theorem. Equivalently, increasing the dimension of the central subspace will be orthogonal to the goal of minimizing information compression.

## 5. Numerical Illustration

The example in this paper comes from a randomized trial of opioid-dependent participants. Opioid addiction involving both heroin and diverted prescription opioid use represents major public health epidemics in the United States [43]. Currently, two treatments that are effective for opioid addition are agonist therapy with either buprenorphine (BUP) or methadone (MET). The study by Saxon et al. [44] was to determine if there were differences between BUP and MET with respect to liver function in subjects being treated for opioid dependence. Subjects who met the study inclusion criteria were randomized to BUP or MET; there was a total of $n = 832$ subjects in the analysis. Here, we will focus on the change in weight from baseline to week 12 as the dependent variable. Predictor variables include weight at baseline, treatment, gender and ethnicity. Assuming that the central subspace is of dimension one, using sliced inverse regression [28], we estimate the basis to be $(-0.38, -0.75, 0.53, 0.03)$. Thus, we would estimate the first direction to be

$$-0.38\text{Tx} - 0.75\text{Gender} + 0.53\text{Ethn} + 0.03\text{BaseWt}. \tag{12}$$

Note that in the classical framework of sufficient dimension reduction, the interpretation of the estimate is problematic. This is due to the fact that treatment, gender and ethnicity are binary variables. This means that viewed as an estimand; the central subspace formally does not exist. Having said this, the framework in this paper would view (12) as the linear combination of the variables that achieves maximum information compression in the predictors while simultaneously minimizing information loss between the covariates with the outcome variable. Note that this interpretation does not require the existence of a central subspace.

Naik and Tsai [45] proposed the use of partial least squares (PLS) as a means of sufficient dimension reduction in the situation where the dimension of the central subspace equals one. For these data, we would estimate a combination analogous to (12) based on partial least squares by

$$0.0001\text{Tx} + 0.0003\text{Gender} - 0.0001\text{Ethn} - 0.0268\text{BaseWt} \tag{13}$$

Comparing the magnitudes of (12) and (13), SIR estimates larger relative weights with the exception of weight at baseline. Again, we can interpret the PLS estimate as the linear combination of the variables that achieves maximum information compression in the covariates while simultaneously minimizing information loss regarding their association with the outcome variable. A Github repository illustrating these analyses can be found at http://github.com/GhoshLab/ITSDR/.

## 6. Discussion

In this article, we have attempted to reinterpret the sufficient dimension reduction methodology in the statistical literature using connections to information theory. This link, and in particular to that of the theory of information bottleneck [4], allows for some new insights and interpretations to occur:

1. We can avoid the goal of SDR as estimating a parameter, namely the basis of the central subspace, and view it instead as a means for information compression while simultaneously preserving association with an outcome variable. This information-theoretic view can allow for one to relax distributional assumptions in a way that is different from the $\sigma-$field approach described in [16].
2. By recognizing that the Gaussian bottleneck information theorem (Theorem 3.1 of [17]) is identical to solving a generalized eigenvalue problem, we can extend the results of [17] to a variety of sufficient dimension reduction methods. There, we see that the goals of information compression and central subspace dimension estimation are dual to each other.

Our hope is that this initial exploration of information theory with sufficient dimension reduction will allow for the adaptation and extension of information theoretic concepts into the SDR literature. We envision there being connections and development of methodologies for SDR in time series [46] and online [47,48] problems. This is currently under investigation.

**Funding:** This research was funded by National Science Foundation, Grant No. DMS 1914739 and National Cancer Institute, Grant No. R01 CA129102.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Acknowledgments:** The author would like to acknowledge support from NIH R01 CA129102 and NSF DMS 1914739. He would like to thank Thao Vu, Elin Shaddox and Debmalya Nandy, who provided useful feedback on the paper. In addition, he would like to acknowledge the two referees, whose feedback improved the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Li, B. *Sufficient Dimension Reduction: Methods and Applications with R*; CRC Press: Boca Raton, FL, USA, 2018.
2. Brillinger, D.R. A generalized linear model with "Gaussian" regressor variables. In *Selected Works of David Brillinger*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 589–606.
3. Li, K.C.; Duan, N. Regression analysis under link violation. *Ann. Stat.* **1989**, *17*, 1009–1052. [CrossRef]
4. Tishby, N.; Pereira, F.; Bialek, W.; Hajek, B.; Sreenivas, R. The informational bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, Florham Park, NJ, USA, 30 September 1999. Available online: https://www.bibsonomy.org/bibtex/15bd5efbf394791da00b09839b9a5757 (accessed on 11 December 2021).
5. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473. [CrossRef]
6. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 14–20. [CrossRef]
7. Slonim, N.; Tishby, N. *Agglomerative Information Bottleneck*; ACM: New York, NY, USA, 1999; Volume 4.
8. Slonim, N.; Tishby, N. Document clustering using word clusters via the information bottleneck method. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 24–28 July 2000; pp. 208–215.
9. Slonim, N.; Friedman, N.; Tishby, N. Multivariate information bottleneck. *Neural Comput.* **2006**, *18*, 1739–1789. [CrossRef]
10. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015; pp. 1–5.
11. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. *J. Stat. Mech. Theory Exp.* **2019**, *2019*, 124020. [CrossRef]
12. Xia, Y.; Tong, H.; Li, W.K.; Zhu, L.X. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B* **2002**, *64*, 299–346. [CrossRef]
13. Fukumizu, K.; Bach, F.R.; Jordan, M.I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **2004**, *5*, 73–99.
14. Fukumizu, K.; Bach, F.R.; Jordan, M.I. Kernel dimension reduction in regression. *Ann. Stat.* **2009**, *37*, 1871–1905. [CrossRef]
15. Li, B.; Artemiou, A.; Li, L. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Stat.* **2011**, *39*, 3182–3210. [CrossRef]
16. Lee, K.Y.; Li, B.; Chiaromonte, F. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann. Stat.* **2013**, *41*, 221–249. [CrossRef]

17. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information Bottleneck for Gaussian Variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
18. Wang, Q.; Yin, X.; Critchley, F. Dimension reduction based on the Hellinger integral. *Biometrika* **2015**, *102*, 95–106. [CrossRef]
19. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412. [CrossRef]
20. Yin, X. Canonical correlation analysis based on information theory. *J. Multivar. Anal.* **2004**, *91*, 161–176. [CrossRef]
21. Iaci, R.; Yin, X.; Sriram, T.; Klingenberg, C.P. An informational measure of association and dimension reduction for multiple sets and groups with applications in morphometric analysis. *J. Am. Stat. Assoc.* **2008**, *103*, 1166–1176. [CrossRef]
22. Yin, X.; Sriram, T. Common canonical variates for independent groups using information theory. *Stat. Sin.* **2008**, *18*, 335–353.
23. Xue, Y.; Wang, Q.; Yin, X. A unified approach to sufficient dimension reduction. *J. Stat. Plan. Inference* **2018**, *197*, 168–179. [CrossRef]
24. Cook, R.D.; Ni, L. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Am. Stat. Assoc.* **2005**, *100*, 410–428. [CrossRef]
25. Yao, W.; Nandy, D.; Lindsay, B.G.; Chiaromonte, F. Covariate information matrix for sufficient dimension reduction. *J. Am. Stat. Assoc.* **2019**, *114*, 1752–1764. [CrossRef]
26. Lauritzen, S.L. *Graphical Models*; Clarendon Press: Oxford, UK, 1996; Volume 17.
27. Ichimura, H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econom.* **1993**, *58*, 71–120. [CrossRef]
28. Li, K.C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **1991**, *86*, 316–327. [CrossRef]
29. Cook, R.D. *Regression Graphics: Ideas for Studying Regressions through Graphics*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 482.
30. Yin, X.; Li, B.; Cook, R.D. Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivar. Anal.* **2008**, *99*, 1733–1757. [CrossRef]
31. Li, K.C. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Am. Stat. Assoc.* **1992**, *87*, 1025–1039. [CrossRef]
32. Wu, Q.; Liang, F.; Mukherjee, S. Kernel sliced inverse regression: Regularization and consistency. In *Abstract and Applied Analysis*; Hindawi: London, UK, 2013; Volume 2013.
33. Hall, P.; Li, K.C. On almost linearity of low dimensional projections from high dimensional data. *Ann. Stat.* **1993**, *21*, 867–889. [CrossRef]
34. Chiaromonte, F.; Cook, R.D.; Li, B. Sufficient dimension reduction in regressions with categorical predictors. *Ann. Stat.* **2002**, *30*, 475–497. [CrossRef]
35. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
36. Berge, C. *Hypergraphs: Combinatorics of Finite Sets*; Elsevier: Amsterdam, The Netherlands, 1984; Volume 45.
37. Cover, T.M.; Thomas, J. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.
38. Berk, R.; Buja, A.; Brown, L.; George, E.; Kuchibhotla, A.K.; Su, W.; Zhao, L. Assumption lean regression. *Am. Stat.* **2019**, *75*, 76–84. [CrossRef]
39. Ye, Z.; Weiss, R.E. Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Stat. Assoc.* **2003**, *98*, 968–979. [CrossRef]
40. Luo, W.; Li, B. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **2016**, *103*, 875–887. [CrossRef]
41. Luo, W.; Li, B. On order determination by predictor augmentation. *Biometrika* **2021**, *108*, 557–574. [CrossRef]
42. Barber, R.F.; Candès, E.J. Controlling the false discovery rate via knockoffs. *Ann. Stat.* **2015**, *43*, 2055–2085. [CrossRef]
43. Substance Abuse and Mental Health Services Administration. *Opioid Treatment Program (OTP) Guidance*; Substance Abuse and Mental Health Services Administration: Rckville, MD, USA, 2020.
44. Saxon, A.J.; Ling, W.; Hillhouse, M.; Thomas, C.; Hasson, A.; Ang, A.; Doraimani, G.; Tasissa, G.; Lokhnygina, Y.; Leimberger, J.; et al. Buprenorphine/naloxone and methadone effects on laboratory indices of liver health: A randomized trial. *Drug Alcohol Depend.* **2013**, *128*, 71–76. [CrossRef] [PubMed]
45. Naik, P.; Tsai, C.L. Partial least squares estimator for single-index models. *J. R. Stat. Soc. Ser. B* **2000**, *62*, 763–771. [CrossRef]
46. Li, K.C.; Shedden, K. Identification of shared components in large ensembles of time series using dimension reduction. *J. Am. Stat. Assoc.* **2002**, *97*, 759–765. [CrossRef]
47. Cai, Z.; Li, R.; Zhu, L. Online Sufficient Dimension Reduction Through Sliced Inverse Regression. *J. Mach. Learn. Res.* **2020**, *21*, 1–25.
48. Artemiou, A.; Dong, Y.; Shin, S.J. Real-time sufficient dimension reduction through principal least squares support vector machines. *Pattern Recognit.* **2021**, *112*, 107768. [CrossRef]