Statistica Sinica **32** (2022), 2315-2337
doi:https://doi.org/10.5705/ss.202022.0097

SLICED INVERSE REGRESSION IN METRIC SPACES

Joni Virta, Kuang-Yao Lee and Lexin Li

University of Turku, Temple University and University of California at Berkeley

Abstract: In this article, we propose a general nonlinear sufficient dimension reduction (SDR) framework when both the predictor and the response lie in some general metric spaces. We construct reproducing kernel Hilbert spaces with kernels that are fully determined by the distance functions of the metric spaces, and then leverage the inherent structures of these spaces to define a nonlinear SDR framework. We adapt the classical sliced inverse regression within this framework for the metric space data. Next we build an estimator based on the corresponding linear operators, and show that it recovers the regression information in an unbiased manner. We derive the estimator at both the operator level and under a coordinate system, and establish its convergence rate. Lastly, we illustrate the proposed method using synthetic and real data sets that exhibit non-Euclidean geometry.

Key words and phrases: Covariance operator, metric space, reproducing kernel Hilbert space, sliced inverse regression, sufficient dimension reduction.

1. Introduction

High-dimensional data are now commonplace in almost every branch of science and business, and dimension reduction plays a central role in analyzing such data. A particularly useful reduction paradigm is *sufficient dimension reduction* (SDR), which embodies a family of methods that aim to reduce the dimensionality in a regression setting, without losing any information. Since the pioneering work of the *sliced inverse regression* (Li (1991, SIR)), SDR has developed rapidly. For a univariate response Y and a p-dimensional predictor X, SDR seeks a low-dimensional representation, usually in the form of linear combinations $\beta^\intercal X$, for a $p \times d$ matrix $\beta = (\beta_1, \ldots, \beta_d)$ with $d \leq p$, such that

$$Y \perp\!\!\!\perp X \mid \beta_1^\intercal X, \ldots, \beta_d^\intercal X. \tag{1.1}$$

As such, $\beta^\intercal X$ contains full regression information of Y given X, and the dimension is reduced because d is often much smaller than p. SDR then seeks the minimum subspace spanned by β, called the *central subspace*, which exists uniquely under

Corresponding author: Kuang-Yao Lee, Department of Statistical Science, Temple University, Philadelphia, PA 19122, USA. E-mail: kuang-yao.lee@temple.edu.

very mild conditions (Yin, Li and Cook (2008)). Numerous SDR methods have since been proposed based on SIR (Li (1991)), mostly in a model-free fashion that does not impose a specific parametric form on the association between Y and $\beta^\intercal X$. Examples include the works of Cook and Weisberg (1991), Li (1992), Cook and Li (2002), Xia et al. (2002), Li and Wang (2007), and Ma and Zhu (2012, 2013), among many others. See also Li (2018b) for a comprehensive review.

The SDR in (1.1) achieves a *linear dimension reduction*, because the low-dimensional representation takes the form of linear combinations of X. However, although it preserves the original coordinates of X and is easier to interpret, it is also less flexible. A more recent line of SDR research instead seeks *nonlinear dimension reduction* (Fukumizu, Bach and Jordan (2004, 2009); Li, Artemiou and Li (2011); Lee, Li and Chiaromonte (2013); Li and Song (2017)), such that

$$Y \perp\!\!\!\perp X \mid f_1(X), \ldots, f_d(X), \tag{1.2}$$

where $f_1, \ldots, f_d$ are some functions in a Hilbert space. Nonlinear SDR is more flexible, and may require fewer functions than its linear counterpart to capture the full regression information. However, in general, it is also more difficult to interpret.

Despite the substantial progress of SDR, most existing SDR solutions target data in a Euclidean space. However, modern data objects are becoming increasingly complex, and often reside in non-Euclidean spaces. Such data are routinely collected in applications such as medical imaging, computational biology, and computer vision, and thus it is of great interest to understand the associations between these complex data objects (Lin et al. (2017); Cornea et al. (2017); Dubey and Müller (2019); Petersen and Müller (2019); Lin and Yao (2019); Pan et al. (2020)). As examples, we consider geometric data, positive-definite matrix data, and compositional data. For instance, in brain structural and functional connectivity analyses (Zhu et al. (2009); Zhang, Sun and Li (2020)), the data are usually in the form of positive-definite matrices that measure the connectivity strengths of pairs of nodes of a network and admit a certain manifold structure. In chemistry, geology, and microbiome analysis (Lu, Shi and Li (2019)), the data are the proportions of individual components that sum to a fixed constant. There are many other examples of complex object data (Wang and Marron (2007)). In all these examples, the data reside in some non-Euclidean spaces, and a proper metric is needed to characterize the intrinsic features of the data.

We propose a general nonlinear SDR framework when both the predictor and the response lie in some general, and possibly different, metric spaces. Our key

idea is to construct a pair of reproducing kernel Hilbert spaces (RKHSs), the kernels of which are fully determined by the distance functions of the metric spaces. We then leverage the inherent structures of these spaces to define a nonlinear SDR framework for the metric space data. Here, we adapt the sliced inverse regression of Li (1991) within this framework. We build the estimator based on some linear operators, and show that it recovers the regression information in an unbiased manner. We derive the estimator at both the operator level and under a coordinate system. We also establish the convergence rate of the estimator under both settings when the response lies on a general metric space, and when the response is categorical. We illustrate the proposed method using synthetic and real data sets that exhibit non-Euclidean geometry.

Our proposal is related to, but also differs from the nonlinear SDR method of Lee, Li and Chiaromonte (2013), as well as some recent SDR solutions involving functional or non-Euclidean data, such as those of Yeh, Huang and Lee (2008), Li and Song (2017), Tomassi et al. (2019), Ying and Yu (2020), and Lee and Li (2022). In particular, Lee, Li and Chiaromonte (2013) developed a general framework for nonlinear SDR in which they estimate the functions $f_1, \ldots, f_d$ in (1.2) as the eigenfunctions of some linear operator defined on a Hilbert space $\mathcal{H}$. However, they target Euclidean data, and take $\mathcal{H}$ to be an L_2-space at the population level and an RKHS at the sample level. Our framework is similar to theirs, but we consider data residing in a general metric space. Moreover, we take $\mathcal{H}$ to be an RKHS at both the population and the sample levels, which makes the connection between the population and sample versions of the estimation procedure more transparent. Yeh, Huang and Lee (2008) proposed a kernel SIR under the framework of (1.2), but require a functional version of the linearity condition. We instead adopt a general form of conditional independence based on σ-fields, and avoid relying on the linearity condition. Li and Song (2017) considered nonlinear SDR for functional data, where X is a function residing in some Hilbert space, and Lee and Li (2022) studied linear SDR when X and Y are both functions in some Hilbert space. In contrast, we consider more general data objects than functional data. Tomassi et al. (2019) developed linear SDR for compositional data, but restricted their solution to a specific set of parametric models for the conditional distribution of X given Y. Ying and Yu (2020) developed SDR when the response is in a metric space and the predictors reside in a Euclidean space. However, because the dimension reduction is performed for the predictors, our method differs considerably from that of Ying and Yu (2020).

The rest of the article is organized as follows. In Section 2, we develop the general framework for nonlinear SDR for data in metric spaces, and in Section

3, we derive the metric version of SIR under this framework. In Section 4, we describe a finite-sample implementation, and in Section 5, we study the convergence properties of the estimator. In Section 6, we present the results of our numerical studies. All proofs are provided in the online Supplementary Material.

2. Nonlinear SDR for Metric Space Data

In this section, we propose a general framework for conducting nonlinear SDR on data residing in arbitrary metric spaces. First, we define a minimal σ-field that captures the full regression information. Then, we construct RKHSs for X and Y from the metric spaces, and use these RKHSs to define a representation of the minimal σ-field that is easier to estimate.

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space. Let (Ω_X^0, d_X) and (Ω_Y^0, d_Y) be arbitrary separable metric spaces in which the predictor and the response, respectively, take values. We make no further assumption on the data space and, depending on Ω_X^0 and Ω_Y^0, there may be multiple feasible choices for the metrics d_X and d_Y. For instance, in Section 6, we take Ω_X^0 to be some manifold spaces and consider different choices of metrics for d_X.

Let $\mathcal{F}_X$ and $\mathcal{F}_Y$ be the Borel σ-fields generated by the open sets in the metric topologies in Ω_X^0 and Ω_Y^0, respectively. Consider $X : \Omega \to \Omega_X^0$ to be an $\mathcal{F}/\mathcal{F}_X$-measurable random variable with the distribution $P_X = P \circ X^{-1}$, and $Y : \Omega \to \Omega_Y^0$ to be an $\mathcal{F}/\mathcal{F}_Y$-measurable random variable with the distribution $P_Y = P \circ Y^{-1}$. For simplicity, suppose the joint random variable (X, Y) is $\mathcal{F}/(\mathcal{F}_X \times \mathcal{F}_Y)$-measurable. Let $P_{X|Y} : \mathcal{F}_X \times \Omega_Y^0 \to \mathbb{R}$ be the conditional distribution of X given $Y = y$, and suppose the set $\{P_{X|Y}(\cdot \mid y) \mid y \in \Omega_Y^0\}$ is dominated by a σ-finite measure. Let σ_X be the σ-field generated by X. We adopt the following definition from Lee, Li and Chiaromonte (2013).

Definition 1. A sub-σ-field $\mathcal{G}$ of σ_X is said to be an SDR σ-field for Y given X if the random elements Y and X are conditionally independent given $\mathcal{G}$, in that $Y \perp\!\!\!\perp X \mid \mathcal{G}$. When the set of conditional distributions $\{P_{X|Y}(\cdot \mid y) \mid y \in \Omega_Y^0\}$ is dominated by a σ-finite measure, the intersection of all SDR σ-fields is itself an SDR σ-field, called the central σ-field, and is denoted by $\mathcal{G}_{Y|X}$.

Definition 1 suggests that there exists a unique smallest SDR σ-field. In our pursuit of nonlinear SDR, we seek a set of functions $f_1, \ldots, f_d$, lying in some suitable function space $\mathcal{H}_X$, that are $\mathcal{G}_{Y|X}$-measurable, and achieve the dimension reduction by replacing X with the corresponding sufficient predictors $f_1(X), \ldots, f_d(X)$.

A natural candidate for the function space $\mathcal{H}_X$ is $L_2(P_X)$, the class of all square integrable functions $f : \Omega_X^0 \to \mathbb{R}$; see Lee, Li and Chiaromonte (2013).

We instead take $\mathcal{H}_X$ to be a suitably defined RKHS, which makes the subsequent methodology and theory development considerably simpler. To connect the RKHS $\mathcal{H}_X$ to the metric structure of the space Ω_X^0, we consider a positive semi-definite kernel $\kappa_X : \Omega_X^0 \times \Omega_X^0 \to \mathbb{R}$, for which there exists a function $\rho : \mathbb{R} \to \mathbb{R}$, such that, for all $x_1, x_2 \in \Omega_X^0$,

$$\kappa(x_1, x_2) = \rho\{d_X(x_1, x_2)\}, \tag{2.1}$$

where d_X is the metric of Ω_X^0. We further impose the following finite second-order moment requirement for the kernel function, which is essentially the RKHS-equivalent of requiring a random variable to be square integrable, and is a rather mild condition.

Assumption 1. *Suppose* $\mathrm{E}\{\kappa_X(X, X)\} < \infty$, *and* $\mathrm{E}\{\kappa_Y(Y, Y)\} < \infty$.

There are multiple choices for this type of kernel function, including the Gaussian kernel and the Laplace kernel, among others. Here, we use a Gaussian kernel with a positive covariance.

Given the kernels κ_X and κ_Y, let $\mathcal{H}_X^0$ and $\mathcal{H}_Y^0$ be the RKHSs generated by κ_X and κ_Y, respectively. By Assumption 1, we have that $\mathcal{H}_X^0 \subseteq L_2(P_X)$ and $\mathcal{H}_Y^0 \subseteq L_2(P_Y)$. Moreover, by the Riesz representation theorem, there exist a unique mean element $\mu_X \in \mathcal{H}_X^0$ and a unique covariance operator Σ_{XX}^0, such that

$$\begin{aligned} \langle f, \mu_X \rangle_{\mathcal{H}_X^0} &= \mathrm{E}\{f(X)\}, \quad \text{for all} \quad f \in \mathcal{H}_X^0, \\ \langle f, \Sigma_{XX}^0 f' \rangle_{\mathcal{H}_X^0} &= \mathrm{Cov}\{f(X), f'(X)\}, \quad \text{for all} \quad f, f' \in \mathcal{H}_X^0. \end{aligned}$$

Note that every $f_0 \in \ker(\Sigma_{XX}^0)$ satisfies $\mathrm{Var}\{f_0(X)\} = \langle f_0, \Sigma_{XX}^0 f_0 \rangle_{\mathcal{H}_X^0} = 0$, and is almost surely equal to a constant, where $\ker(\cdot)$ denotes the null space. As such, we restrict our attention to $\mathcal{H}_X = \overline{\mathrm{ran}}(\Sigma_{XX}^0)$, where $\mathrm{ran}(\cdot)$ denotes the range, and $\overline{\mathrm{ran}}(\cdot)$ denotes the closure of the range.

Lemma 1. *Suppose Assumption* 1 *holds. There exists a set* $\Omega_X \subseteq \Omega_X^0$, *such that* $P_X(\Omega_X) = 1$ *and* $\kappa_X(\cdot, x) - \mu_X \in \mathcal{H}_X$, *for all* $x \in \Omega_X$.

Lemma 1 states that the functions $\kappa_X(\cdot, x) - \mu_X$, for $x \in \Omega_X$, belong to the space $\mathcal{H}_X$, which allows us to perform centering using the inner product, $\langle f, \kappa_X(\cdot, x) - \mu_X \rangle_{\mathcal{H}_X} = f(x) - \mathrm{E}\{f(X)\}$. The proof of Lemma 1 also shows that the space $\mathcal{H}_X$ admits an alternative characterization, $\mathcal{H}_X = \overline{\mathrm{span}}\{\kappa_X(\cdot, x) - \mu_X : x \in \Omega_X\}$, where $\overline{\mathrm{span}}(\cdot)$ denotes the closure of the space spanned by the set of functions. A similar result was obtained by Li and Song (2017, Lemma 1). However, their proof

implicitly assumes that the empty set is the only set for which P_X assigns a zero probability, essentially ruling out all continuous distributions; our Lemma 1 fixes this issue. Furthermore, this characterization does not imply that the elements $f \in \mathcal{H}_X$ are centered in the sense that $\mathrm{E}\{f(X)\} = 0$. Instead, focusing on $\mathcal{H}_X$ removes the constant functions that are of no interest to us in terms of dimension reduction. We construct μ_Y, Σ^0_{YY}, and the RKHS $\mathcal{H}_Y$ in an analogous manner.

Definition 2. We call the set of all $f \in \mathcal{H}_X$ that are $\mathcal{G}_{Y|X}$-measurable the *central class*, and denote this set by $\mathcal{S}_{Y|X}$.

We make two remarks. First, our notion of dimension reduction is based on the smallest SDR σ-field, that is, the central σ-field. In our setting, the concept of "dimensionality" is less obvious than that in the classical SDR setting, where it is simply the dimension of the central subspace. This is because there are sets that generate the same σ-field, but with different dimensions. Nevertheless, our formulation is useful when one is interested in reducing the dimensionality in the class sense, because the central class induced by the central σ-field contains all sets of functions that generate the same σ-field, and we seek the smallest one. Second, the relation between the central σ-field $\mathcal{G}_{Y|X}$ and the central class $\mathcal{S}_{Y|X}$ is analogous to the relation between the central subspace and the sufficient predictors in the classical setting. That is, in lieu of estimating $\mathcal{G}_{Y|X}$, we search for subsets of elements of $\mathcal{S}_{Y|X}$, which are more concrete and easier to estimate.

3. Metric SIR

In this section, we derive the population-level SIR for metric space data. In the classical SIR (Li (1991)), X and Y both lie in a Euclidean space, and one estimates the central subspace using the range of the matrix,

$$\mathrm{Var}(X)^{-1}\mathrm{Var}\{\mathrm{E}(X \mid Y)\}. \tag{3.1}$$

We next derive the operator analogue for (3.1) for two cases: the general case of Y residing in a metric space, and the special case of Y being a discrete random variable.

3.1. Metric response

We first define a number of covariance operators that serve as the building blocks of our nonlinear metric SIR procedure:

$$\Sigma_{XX} : \mathcal{H}_X \to \mathcal{H}_X, \langle f, \Sigma_{XX} f' \rangle_{\mathcal{H}_X} = \mathrm{Cov}\{f(X), f'(X)\},$$

$$\Sigma_{XY} : \mathcal{H}_Y \to \mathcal{H}_X,\ \langle f, \Sigma_{XY} g\rangle_{\mathcal{H}_X} = \mathrm{Cov}\{f(X), g(Y)\}, \tag{3.2}$$
$$\Sigma_{YY} : \mathcal{H}_Y \to \mathcal{H}_Y,\ \langle g', \Sigma_{YY} g\rangle_{\mathcal{H}_Y} = \mathrm{Cov}\{g'(Y), g(Y)\},$$

for $f, f' \in \mathcal{H}_X$ and $g, g' \in \mathcal{H}_Y$. In addition, the cross-covariance operator $\Sigma_{YX} : \mathcal{H}_X \to \mathcal{H}_Y$ can be obtained as $\Sigma_{YX} = \Sigma_{XY}^*$, the adjoint of the operator Σ_{XY}. Furthermore, because $\mathcal{H}_X = \overline{\mathrm{ran}}(\Sigma_{XX}^0)$, we have $\ker(\Sigma_{XX}) = \{0\}$ and $\overline{\mathrm{ran}}(\Sigma_{XX}) = \mathcal{H}_X$.

We next introduce two regularity conditions.

Assumption 2. *Suppose that $\mathcal{H}_X + \mathbb{R}$ and $\mathcal{H}_Y + \mathbb{R}$ are dense in $L_2(P_X)$ and $L_2(P_Y)$, respectively, where $+$ denotes the direct sum.*

Assumption 3. *Suppose* $\mathrm{ran}(\Sigma_{YX}) \subseteq \mathrm{ran}(\Sigma_{YY})$ *and* $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX})$.

Assumption 2 is typical in kernel learning and holds in general, for example, when κ_X is a Gaussian kernel (Fukumizu, Bach and Jordan (2009)). In this assumption, by "dense" we mean that, for every $f \in L_2(P_X)$, there exists a sequence of elements $f_n \in \mathcal{H}_X$, such that $\mathrm{var}\{f(X) - f_n(X)\} \to 0$, as $n \to \infty$. Assumption 3 is essentially a smoothness condition on the relation between X and Y (Li (2018a)). Similar conditions are common in the SDR literature (Ying and Yu (2020); Li and Song (2022)). Assumption 3 guarantees that the operator $\Sigma_{YY}^\dagger \Sigma_{YX}$ is both well-defined and bounded (Douglas (1966, Thm. 1)), where $\dagger$ denotes the Moore–Penrose pseudo-inverse of Σ_{YY}; see Li (2018a) for details on the Moore–Penrose pseudo-inverse of an operator.

The next lemma provides some useful expressions for the conditional moments of X given Y at the operator level that are essential in the construction of the operator analogue for the SIR estimator (3.1). In addition, they help turn conditional moments into unconditional moments, thus avoiding the slicing step in the original SIR.

Lemma 2. *Suppose Assumptions* 1, 2, *and* 3 *hold. Then,*

(a) for any $f \in \mathcal{H}_X$, $\mathrm{E}\{f(X) \,|\, Y\} - \mathrm{E}\{f(X)\} = \langle \Sigma_{YY}^\dagger \Sigma_{YX} f, \kappa_Y(\cdot, Y) - \mu_Y\rangle_{\mathcal{H}_Y}$;

(b) for any $f, f' \in \mathcal{H}_X$, $\mathrm{Cov}[\mathrm{E}\{f(X) \,|\, Y\}, \mathrm{E}\{f'(X) \,|\, Y\}] = \langle f, \Sigma_{XY}\Sigma_{YY}^\dagger \Sigma_{YX} f'\rangle_{\mathcal{H}_X}$.

By Lemma 2, the operator $\Sigma_{XY}\Sigma_{YY}^\dagger\Sigma_{YX}$ can be viewed as the analogue of the matrix $\mathrm{Var}\{\mathrm{E}(X \mid Y)\}$ in (3.1), and the operator $\Sigma_{XX}^\dagger$ can be viewed as the analogue of $\mathrm{Var}(X)^{-1}$ in (3.1). Consequently, a direct operator counterpart of (3.1) is

$$\Lambda_{\mathrm{SIR}} = \Sigma_{XX}^\dagger \Sigma_{XY} \Sigma_{YY}^\dagger \Sigma_{YX}. \tag{3.3}$$

This operator is well-defined, by Assumption 3. Moreover, if we choose linear kernels κ_X and κ_Y, then Λ_{SIR} reduces precisely to the matrix of the canonical correlation analysis (CCA).

The next theorem shows that the operator Λ_{SIR} is bounded and that the closure of its range is unbiased for the central class, mirroring the classical SIR for linear SDR of Euclidean data. We need an additional regularity condition.

Assumption 4. *Suppose the set* $\mathrm{ran}(\Sigma_{XX}) \cap \mathcal{S}_{Y|X}^{\perp}$ *is dense in the set* $\mathcal{S}_{Y|X}^{\perp}$*, where the orthogonal complement is taken with respect to* $\mathcal{H}_X$*.*

Assumption 4 requires that the intersection between $\mathrm{ran}(\Sigma_{XX})$ and $\mathcal{S}_{Y|X}^{\perp}$ is suitably rich in $\mathcal{S}_{Y|X}^{\perp}$. This is a mild condition, because $\mathrm{ran}(\Sigma_{XX})$ is, by definition, dense in its closure $\mathcal{H}_X$. A similar condition is imposed implicitly in Li and Song (2017).

Theorem 1. *Suppose Assumptions* 1 *to* 4 *hold. Then,* Λ_{SIR} *is a bounded operator and* $\overline{\mathrm{ran}}(\Lambda_{\mathrm{SIR}}) \subseteq \mathcal{S}_{Y|X}$*.*

Theorem 1 suggests that we can recover the central class using the range of Λ_{SIR}, or equivalently, by using the spectral decomposition of $\Lambda_{\mathrm{SIR}}\Lambda_{\mathrm{SIR}}^{*}$. This is the foundation of our estimation procedure, developed in Section 4. We call our proposed nonlinear SDR method based on Λ_{SIR} the *metric sliced inverse regression* (MSIR).

3.2. Discrete response

Next, we consider a special case in which Y lies in the usual Euclidean space and is discrete. This scenario is perhaps most often encountered in real applications. The main difference between this special case and the general case is that, when Y is discrete, we can obtain direct RKHS representations for the conditional moments, rather than resorting to unconditional representations, as in Lemma 2.

Specifically, suppose $\Omega_Y^0 = \{1, \ldots, K\}$, and let $\pi_k = P(Y = k)$ and $\pi_k > 0$, for all $k \in \Omega_Y^0$. By the Riesz representation theorem, elements $\gamma_{X|k} \in \mathcal{H}_X$, for $k = 1, \ldots, K$, exist such that, for any $f \in \mathcal{H}_X$,

$$\mathrm{E}\{f(X) \mid Y = k\} - \mathrm{E}\{f(X)\} = \langle \gamma_{X|k}, f \rangle_{\mathcal{H}_X}.$$

The elements $\gamma_{X|k}$ provide a discrete counterpart of Lemma 2(a). We then define the covariance operator

$$\Gamma_{XX|Y} = \sum_{k=1}^{K} \pi_k (\gamma_{X|k} \otimes \gamma_{X|k}) : \mathcal{H}_X \to \mathcal{H}_X, \tag{3.4}$$

where $\otimes$ denotes the tensor product. Note that, for any $f, f' \in \mathcal{H}_X$,

$$\mathrm{Cov}[\mathrm{E}\{f(X) \mid Y\}, \mathrm{E}\{f'(X) \mid Y\}] = \langle f, \Gamma_{XX|Y} f' \rangle_{\mathcal{H}_X}.$$

Consequently, the counterpart of Λ_{SIR} in (3.3) when Y is categorical is

$$\Lambda_{\mathrm{SIR,D}} = \Sigma_{XX}^{\dagger} \Gamma_{XX|Y}. \tag{3.5}$$

This operator is well-defined under the following smoothness condition, and the closure of its range provides an unbiased estimator of the central class.

Assumption 5. *Suppose* $\mathrm{ran}(\Gamma_{XX|Y}) \subseteq \mathrm{ran}(\Sigma_{XX})$.

Theorem 2. *Suppose Assumptions* 1, 2, 4 *and* 5 *hold. Then,* $\Lambda_{\mathrm{SIR,D}}$ *is a bounded operator and* $\overline{\mathrm{ran}}(\Lambda_{\mathrm{SIR,D}}) \subseteq \mathcal{S}_{Y|X}$.

4. Sample Estimation

In this section, we develop the sample estimator for the proposed metric SIR, first at the operator level, then under a coordinate system, given the independent and identically distributed (i.i.d.) random sample observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of (X, Y).

4.1. Estimation at the operator level

For the general case when the response Y resides in a metric space, we first obtain the sample estimators of the mean elements as $\hat{\mu}_X = \mathrm{E}_n\{\kappa_X(\cdot, X)\}$ and $\hat{\mu}_Y = \mathrm{E}_n\{\kappa_Y(\cdot, Y)\}$, where E_n is the sample mean operator, such that $\mathrm{E}_n \omega = n^{-1} \sum_{i=1}^{n} \omega_i$ for the samples $\omega_1, \ldots, \omega_n$ from ω. We next obtain the sample estimators of the covariance operators Σ_{XX}, Σ_{XY}, and Σ_{YY} in (3.2) as

$$\begin{aligned}
\hat{\Sigma}_{XX} &= \mathrm{E}_n[\{\kappa_X(\cdot, X) - \hat{\mu}_X\} \otimes \{\kappa_X(\cdot, X) - \hat{\mu}_X\}], \\
\hat{\Sigma}_{XY} &= \mathrm{E}_n[\{\kappa_X(\cdot, X) - \hat{\mu}_X\} \otimes \{\kappa_Y(\cdot, Y) - \hat{\mu}_Y\}], \\
\hat{\Sigma}_{YY} &= \mathrm{E}_n[\{\kappa_Y(\cdot, Y) - \hat{\mu}_Y\} \otimes \{\kappa_Y(\cdot, Y) - \hat{\mu}_Y\}].
\end{aligned}$$

Moreover, we have $\hat{\Sigma}_{YX} = \hat{\Sigma}_{XY}^{*}$. We then obtain the sample estimator of the metric SIR operator Λ_{SIR} in (3.3) as

$$\hat{\Lambda}_{\mathrm{SIR}} = (\hat{\Sigma}_{XX} + \tau_1 I)^{-1} \hat{\Sigma}_{XY} (\hat{\Sigma}_{YY} + \tau_2 I)^{-1} \hat{\Sigma}_{YX},$$

where we use a ridge regularization to estimate the pseudo-inverses, τ_1 and τ_2 are the ridge parameters, and I is the identity operator. Finally, we estimate the range of Λ_{SIR} using the spectral decomposition of the operator $\hat{\Lambda}_{\text{SIR}}\hat{\Lambda}^*_{\text{SIR}}$. Suppose $\hat{f}_1, \ldots, \hat{f}_d$ are the d leading eigenfunctions of $\hat{\Lambda}_{\text{SIR}}\hat{\Lambda}^*_{\text{SIR}}$. Then, the estimated sufficient predictors corresponding to the observation $X \in \Omega^0_X$ are $\hat{f}_1(X), \ldots, \hat{f}_d(X)$.

For the special case when Y resides in the usual Euclidean space and is discrete, we obtain the sample estimator of the covariance operator $\Gamma_{XX|Y}$ in (3.4) as

$$\hat{\Gamma}_{XX|Y} = \frac{1}{n}\sum_{k=1}^{K} n_k(\hat{\gamma}_{X|k} \otimes \hat{\gamma}_{X|k}),$$

where n_k is the number of samples belonging to the class k, $\mathbb{I}(\cdot)$ is the indicator function, and $\hat{\gamma}_{X|k} = (n/n_k)\mathrm{E}_n\{\mathbb{I}(Y = k)\kappa_X(\cdot, X)\} - \hat{\mu}_X$, for $k = 1, \ldots, K$. We then obtain the sample estimator of the metric SIR operator $\Lambda_{\text{SIR,D}}$ in (3.5) as

$$\hat{\Lambda}_{\text{SIR,D}} = (\hat{\Sigma}_{XX} + \tau_1 I)^{-1}\hat{\Gamma}_{XX|Y}.$$

Finally, we estimate the range of $\Lambda_{\text{SIR,D}}$ using the spectral decomposition of $\hat{\Lambda}_{\text{SIR,D}}\hat{\Lambda}^*_{\text{SIR,D}}$.

4.2. Estimation under a coordinate representation

We next develop an estimation procedure under a chosen coordinate system. We divide the procedure into three main steps, and focus on the general case in which Y resides in a metric space. We do also briefly discuss the special case in which Y is discrete.

In Step 1, we choose the kernel functions κ_X and κ_Y. Although there are multiple choices of kernel functions, we use the Gaussian kernel throughout. We use the leave-one-out cross-validation method to determine the bandwidth parameters in κ_X and κ_Y, following a similar strategy to that in Lee, Li and Chiaromonte (2013). We then compute the Gram matrices $K_X = (\kappa_X(X_i, X_{i'}))_{i,i'=1}^n \in \mathbb{R}^{n\times n}$ and $K_Y = (\kappa_Y(Y_i, Y_{i'}))_{i,i'=1}^n \in \mathbb{R}^{n\times n}$, where the kernel functions κ_X and κ_Y are evaluated under the given metrics d_X and d_Y, as in (2.1). Let $Q = I - n^{-1}11^\intercal$ denote the centering matrix, where $1 \in \mathbb{R}^n$ is a vector of ones. We then compute the centered version of the Gram matrices as

$$G_X = QK_XQ, \quad \text{and} \quad G_Y = QK_YQ. \tag{4.1}$$

In Step 2, we compute the coordinate representation of the sample metric SIR operator $\hat{\Lambda}_{\text{SIR}}$. Consider the sample counterpart of the space $\mathcal{H}^0_X$, which is the

span of the sample elements, $\hat{\mathcal{H}}_X^0 = \operatorname{span}\{\kappa_X(\cdot, X_i) \mid i = 1, \ldots, n\}$. We impose the following linear independence assumption, which is a mild requirement. When it does not hold, we can simply delete a subset of the elements to obtain a linearly independent set. Alternatively, we can construct a linearly independent basis using the Karhunen–Loève expansion; see, for example, Lee and Li (2022).

Assumption 6. *The elements $\kappa_X(\cdot, X_i)$, for $i = 1, \ldots, n$, are linearly independent.*

Under Assumption 6, the elements $\kappa_X(\cdot, X_i)$, for $i = 1, \ldots, n$, form a basis for $\hat{\mathcal{H}}_X^0$ and, given an arbitrary member $f \in \hat{\mathcal{H}}_X^0$, we define its coordinate $[f] \in \mathbb{R}^n$ as the vector of its coefficients under this basis. As such, for any $f \in \hat{\mathcal{H}}_X^0$ and $X \in \Omega_X^0$, $f(X) = [f]^\intercal k_X(X)$, where $k_X(X) = (\kappa_X(X, X_1), \ldots, \kappa_X(X, X_n))^\intercal$. In addition, we take the inner product of $\hat{\mathcal{H}}_X^0$ to be the bilinear form $(f, f') \mapsto \langle f, f' \rangle_{\hat{\mathcal{H}}_X^0} = [f]^\intercal K_X [f']$, for $f, f' \in \hat{\mathcal{H}}_X^0$, and the Gram matrix K_X is ensured to be positive definite by Assumption 6. Analogously, consider the sample counterpart of the space $\mathcal{H}_X$, which is the span of the centered sample elements, $\hat{\mathcal{H}}_X = \operatorname{span}\{\kappa_X(\cdot, X_i) - \hat{\mu}_X \mid i = 1, \ldots, n\}$. We construct the sample spaces $\hat{\mathcal{H}}_Y^0$ and $\hat{\mathcal{H}}_Y$ similarly.

Correspondingly, following Fukumizu, Bach and Jordan (2009), the coordinates of the sample covariance operators $\hat{\Sigma}_{XX}$, $\hat{\Sigma}_{XY}$, $\hat{\Sigma}_{YX}$, and $\hat{\Sigma}_{YY}$ are

$$[\hat{\Sigma}_{XX}] = n^{-1} G_X, \quad [\hat{\Sigma}_{XY}] = n^{-1} G_Y, \quad [\hat{\Sigma}_{YX}] = n^{-1} G_X, \quad \text{and} \quad [\hat{\Sigma}_{YY}] = n^{-1} G_Y,$$

respectively, where G_X and G_Y are as defined in (4.1). Although this coordinate representation seems to suggest that $\hat{\Sigma}_{YX}$ does not depend on Y, this is not the case. Actually, $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YX}$ share the same coordinate, namely $n^{-1} G_X$, but they involve different sets of bases, because they have different range spaces. For simplicity, we drop the underlying bases in the coordinate bracket notation. However, note that $\hat{\Sigma}_{YX}$ depends on Y through the underlying bases; a similar discussion applies to $\hat{\Sigma}_{XY}$.

We then obtain the coordinate representation of $\hat{\Lambda}_{\mathrm{SIR}}$ in the next lemma. The proof follows immediately by the definition of $\hat{\Lambda}_{\mathrm{SIR}}$, and is thus omitted.

Lemma 3. *The metric SIR operator $\hat{\Lambda}_{\mathrm{SIR}}$ has the coordinate representation*

$$[\hat{\Lambda}_{\mathrm{SIR}}] = G_X^\dagger G_Y G_Y^\dagger G_X, \tag{4.2}$$

where † denotes the Moore–Penrose pseudo-inverse of a matrix.

To improve the numerical stability, we replace the pseudo-inverse $G_X^\dagger$ in Lemma 3 with its ridge-regularized counterpart $\{G_X + \tau_1 I_n\}^{-1}$, where τ_1 is taken

to be $c \times \phi_1(G_X)$, $\phi_1(\cdot)$ is the largest eigenvalue of the designated matrix, and $c = 0.2$. A similar procedure was also employed in Lee and Li (2022). Similarly, we replace $G_Y^\dagger$ with $\{G_Y + \tau_2 I_n\}^{-1}$, where $\tau_2 = c \times \phi_1(G_Y)$.

In Step 3, we estimate the range of $\hat{\Lambda}_{\mathrm{SIR}}$ using the eigen-decomposition of its coordinate in (4.2). Letting $v_1, \ldots, v_d$ denote the d leading eigenvectors of $[\hat{\Lambda}_{\mathrm{SIR}}][\hat{\Lambda}_{\mathrm{SIR}}]^\intercal$, the estimated sufficient predictors corresponding to an observation $X \in \Omega_X^0$ are $v_1^\intercal Q k_X(X), \ldots v_d^\intercal Q k_X(X)$, where $k_X(X) = (\kappa_X(X, X_1), \ldots, \kappa_X(X, X_n))^\intercal$. Alternatively, one can also use the eigenvectors of the matrix $[\hat{\Lambda}_{\mathrm{SIR}}]$.

The computational complexity of our proposed method is $\mathcal{O}(n^3)$. When the sample size n is huge, the computation can be intensive. For such a case, we propose an alternative estimation strategy similar to that of Hung and Huang (2019). That is, we first divide all sample observations into Q disjoint subsets, $\mathcal{I}_1, \ldots, \mathcal{I}_Q$. We then estimate the sufficient predictors, given each subset $\mathcal{I}_q$, for $q = 1, \ldots, Q$. To accommodate possible discrepancies in the signs of the resulting eigenvectors, we choose their signs such that, for each $j = 1, \ldots, d$, we maximize the sum $\sum_{q,q'=1}^{Q} v_{j,q}^\intercal v_{j,q'}$, where $v_{j,q}$ is the jth eigenvector of $[\hat{\Lambda}_{\mathrm{SIR}}][\hat{\Lambda}_{\mathrm{SIR}}]^\intercal$ computed based on the qth subset $\mathcal{I}_q$. We then average the estimated sufficient predictors over all Q subsets to produce the final estimate for the full sample.

For the special case in which Y resides in the usual Euclidean space and is discrete, the coordinate representation of $\gamma_{X|k}$ is $[\hat{\gamma}_{X|k}] = (1/n_k)1_k - (1/n)1$, where the ith element of the vector $1_k \in \mathbb{R}^n$ is the indicator $\mathbb{I}(Y_i = k)$, for $i = 1, \ldots, n$. Correspondingly, the coordinate representation of $\hat{\Lambda}_{\mathrm{SIR,D}}$ is

$$[\hat{\Lambda}_{\mathrm{SIR,D}}] = G_X^\dagger Q \left(\sum_{k=1}^{K} \frac{1}{n_k} 1_k 1_k^\intercal \right) Q G_X.$$

Finally, we briefly comment on the problem of selecting the reduced dimension d in the SDR. Several information criterion-based selection methods have been proposed for the SDR of Euclidean data (Zhu, Miao and Peng (2006); Luo et al. (2009); Xia, Xu and Zhu (2015)). We expect a similar information criterion is applicable for our metric SIR as well, which we leave as a topic for future research.

5. Asymptotic Theory

In this section, we establish the convergence rate of the proposed metric SIR estimator at the operator level for the general Y and categorical Y settings.

We begin with some regularity conditions.

Assumption 7. *Suppose the kernel functions* κ_X *and* κ_Y *are continuous.*

Assumption 8. *Suppose* $\mathrm{E}\{\kappa_X(X,X)^2\} < \infty$ *and* $\mathrm{E}\{\kappa_Y(Y,Y)^2\} < \infty$.

Assumption 9. *Suppose* $\mathrm{ran}(\Sigma_{YX}) \subseteq \mathrm{ran}(\Sigma_{YY}^2)$ *and* $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX}^2)$.

Assumption 7 is quite mild and, together with the separability of the metric spaces Ω_X^0 and Ω_Y^0, ensures that the RKHSs $\mathcal{H}_X$ and $\mathcal{H}_Y$ are separable (Hein and Bousquet (2004)), which, in turn, ensures that $\mathcal{H}_X$ and $\mathcal{H}_Y$ admit countable orthonormal bases. Assumption 8 is analogous to the requirement that a random variable has a finite fourth moment, and is reasonable. Assumption 9 can be viewed as a stronger version of Assumption 3; that is, compared with Assumption 3, the mapping of Σ_{XY} needs to concentrate even more on the leading eigenspaces of Σ_{XX} and Σ_{YY}. This, again, can be understood as a smoothness condition.

In our sample estimation, we use the ridge regularization for the pseudo-inverses. For simplicity, in our theoretical analysis, we suppose the ridge parameters $\tau_1 = \tau_2 = \tau$, and that τ approaches zero as the sample size n diverges. Denote the operator norm of a linear operator $A : \mathcal{H} \to \mathcal{H}'$ as $\|A\|_{\mathrm{OP}} = \sup\{\|Af\|_{\mathcal{H}'} : \|f\|_{\mathcal{H}} = 1\}$. The next theorem establishes the convergence of $\hat{\Lambda}_{\mathrm{SIR}}$ in terms of the operator norm for the general response case.

Theorem 3. *Suppose Assumptions* 7 *to* 9 *hold. Then, as* $n \to \infty$,

$$\left\|\hat{\Lambda}_{\mathrm{SIR}} - \Lambda_{\mathrm{SIR}}\right\|_{\mathrm{OP}} = \mathcal{O}_p\left(\tau + \frac{1}{\tau\sqrt{n}}\right).$$

For the special case of Y being categorical, we replace the smoothness condition of Assumption 9 with the following counterpart.

Assumption 10. *Suppose* $\mathrm{ran}(\Gamma_{XX|Y}) \subseteq \mathrm{ran}(\Sigma_{XX}^2)$.

Theorem 4. *Suppose Assumptions* 7, 8, *and* 10 *hold. Then, as* $n \to \infty$,

$$\left\|\hat{\Lambda}_{\mathrm{SIR,D}} - \Lambda_{\mathrm{SIR,D}}\right\|_{\mathrm{OP}} = \mathcal{O}_p\left(\tau + \frac{1}{\tau\sqrt{n}}\right).$$

Theorems 3 and 4 suggest that our metric SIR estimator is consistent. Its convergence rate consists of two parts. The first part is due to the ridge regularization, and the second part represents the convergence of the sample operators to their population counterparts. If $\tau = n^{-\beta}$, for some constant $\beta > 0$, then the convergence rate becomes $n^{-\beta} + n^{\beta-1/2}$, implying that the best possible convergence rate given by our result is $\mathcal{O}(n^{-1/4})$, achieved when $\beta = 1/4$. Note that this is the same as the rate obtained by Li and Song (2017) in nonlinear SDR for functional data.

6. Numerical Studies

In this section, we investigate the empirical performance of our proposed MSIR, under various choices of distance metrics, and compare it with that of the nonlinear SIR method of Lee, Li and Chiaromonte (2013, GSIR). Although the GSIR method was originally formulated using Euclidean geometry, it can be extended easily to incorporate an arbitrary distance metric.

6.1. Torus manifold data

As the first example, we consider a two-dimensional torus as the predictor, and simulate the response using different distance metrics. A torus is best visualized as a unit square $[0,1]^2$ in which the opposite edges have been "glued together." We consider two generative models:

$$\text{Model 1: } Y_i = d_G\{X_i, (0.5, 0.5)^\intercal\} + \varepsilon_i;$$

$$\text{Model 2: } Y_i = d_G\{X_i, (1, 1)^\intercal\} + \varepsilon_i,$$

where the two-dimensional predictor X_i is uniformly distributed in $[0,1]^2$, the error term ε_i is drawn from a normal distribution with mean zero and variance σ^2, and d_G denotes the geodesic distance. Because the point $(0.5, 0.5)^\intercal$ lies in the middle of the unit square, we have $d_G\{X_i, (0.5, 0.5)^\intercal\} = d_E\{X_i, (0.5, 0.5)^\intercal\}$, where d_E denotes the Euclidean distance. Consequently, in Model 1, the true relation between the response and the predictor is a smooth function of the Euclidean distance between the predictor and the center point of the square, and we expect the two distance functions to perform similarly under Model 1. However, the same is not true for Model 2, where the reference point $(1, 1)^\intercal$ lies at the corner of the square. In this case, the true regression relationship is not a smooth function of the Euclidean distance, but is so for the geodesic distance, making the geodesic distance more favorable under Model 2. For both models, we consider two sample sizes $n = 250, 500$, and two noise levels $\sigma = 0.05, 0.10$. We further divide the data into 80% training samples, and 20% testing samples. We consider two distance metrics, namely, the geodesic distance and the Euclidean distance.

Table 1 reports the distance correlation between the response and the first two estimated sufficient predictors evaluated on the testing samples, averaged over 200 data replications. The results show that the proposed MSIR outperforms the competing GSIR by achieving a higher distance correlation and a smaller standard error. Moreover, the Euclidean metric is slightly better suited to Model 1, where the toroidal geometry plays no role, whereas the geodesic metric is considerably better for Model 2, where the toroidal geometry plays a crucial role. An increased

Table 1. The torus data example: the average distance correlation (with the standard deviation shown in parentheses) between the response and the estimated sufficient predictors.

Model 1	$n = 250$		$n = 500$	
	$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.05$	$\sigma = 0.10$
MSIR d_G	0.912 (0.025)	0.766 (0.058)	0.911 (0.018)	0.777 (0.038)
GSIR d_G	0.719 (0.082)	0.611 (0.088)	0.715 (0.071)	0.599 (0.081)
MSIR d_E	0.926 (0.021)	0.779 (0.060)	0.926 (0.014)	0.790 (0.037)
GSIR d_E	0.654 (0.092)	0.563 (0.091)	0.646 (0.083)	0.552 (0.082)
Model 2	$n = 250$		$n = 500$	
	$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.05$	$\sigma = 0.10$
MSIR d_G	0.912 (0.025)	0.784 (0.054)	0.913 (0.017)	0.775 (0.040)
GSIR d_G	0.726 (0.079)	0.623 (0.094)	0.724 (0.073)	0.616 (0.082)
MSIR d_E	0.841 (0.046)	0.729 (0.067)	0.845 (0.032)	0.722 (0.046)
GSIR d_E	0.602 (0.087)	0.526 (0.098)	0.587 (0.084)	0.509 (0.085)

sample size helps to reduce the standard error of the estimator. Figure 1 provides a visualization of the estimated sufficient predictors for a single data replication under Model 2 with $n = 500$ and $\sigma = 0.05$. The results agree with the qualitative patterns observed in Table 1 that the MSIR produces sufficient predictors that are more informative than those of the GSIR.

6.2. Positive-definite matrix data

As the second example, we consider a positive-definite matrix data example from a neuroimaging-based autism study (Di Martino et al. (2014)). Autism is an increasingly prevalent neurodevelopmental disorder, characterized by symptoms such as social difficulties, communication deficits, stereotyped behaviors, and cognitive delays (Rudie et al. (2013)). The data set consists of $n = 795$ subjects, among whom 362 were diagnosed with autism, and the rest were healthy controls. For each subject, a resting-state functional magnetic resonance imaging (fMRI) scan was obtained, which measures the intrinsic functional architecture of the brain using the correlated synchronizations of brain systems. The corresponding brain functional connectivity network has been shown to alter under different disorders or during different brain developmental stages. Such alterations contain crucial insights for both disorder pathology and the development of the brain (Fox and Greicius (2010)). Therefore, there is great scientific importance in understanding the association between the autism status and the brain connectivity network. Thus, our goal is to produce sufficient predictors that correctly separate the autism patients from the healthy controls.

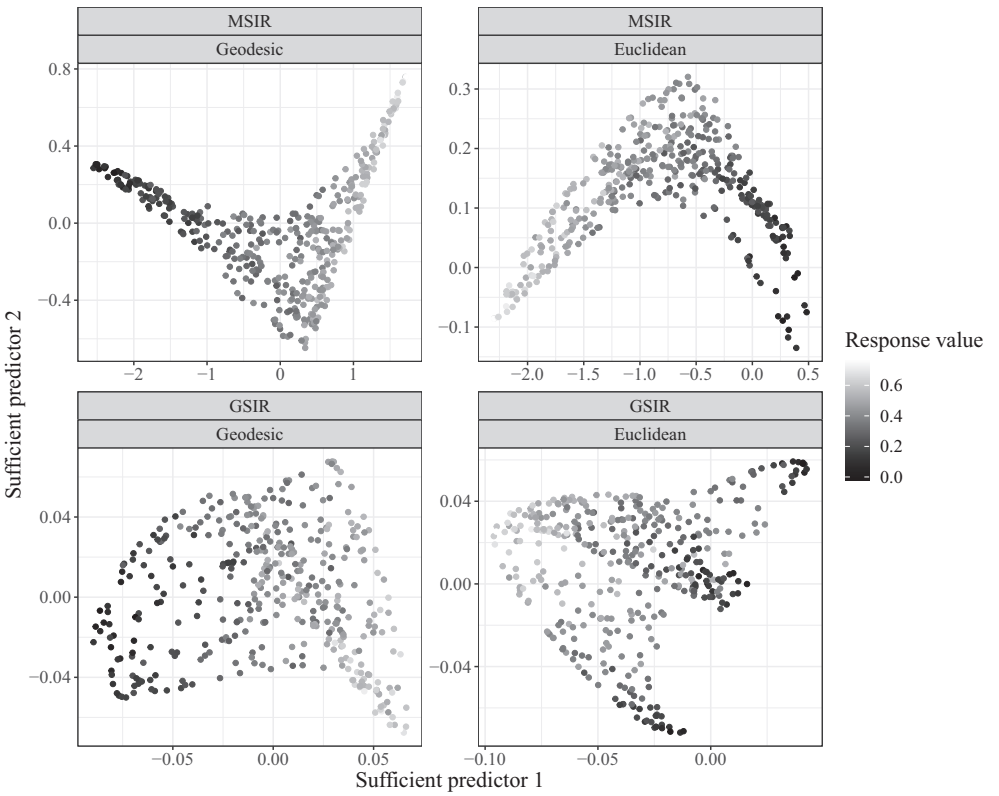

Figure 1. The torus data example: the sufficient predictors under two SDR methods and two distance metrics.

We follow the data-processing procedure of (Sun and Li (2017)), and summarize the brain connectivity network for each subject as a 116×116 correlation matrix, corresponding to the synchronizations of 116 brain regions-of-interest under the commonly used Anatomical Automatic Labeling atlas (Tzourio-Mazoyer et al. (2002)). Moreover, most of the observed connectivity matrices of this data are numerically rank-deficit, with the typical numerical rank ranging from 60 to 80. As such, we employ a common principal components analysis, and project the connectivity matrices onto the space of the top 30 common principal components, such that the minimal eigenvalue is at least 10^{-4} for each resulting matrix.

We consider six distance metrics between two positive-definite matrices, M_1 and M_2: the affine invariant metric, $d_A(M_1, M_2) = \|\mathrm{Log}(M_1^{-1/2} M_2 M_1^{-1/2})\|_F$, where $\mathrm{Log}(\cdot)$ denotes the matrix logarithm, and $\|\cdot\|_F$ denotes the Frobenius norm; the log-Euclidean metric, $d_{LE}(M_1, M_2) = \|\mathrm{Log}(M_1) - \mathrm{Log}(M_2)\|_F$; the S-divergence (Sra (2016)), $d_S(M_1, M_2) = \log|(M_1 + M_2)/2| - (1/2)\log|M_1 M_2|$,

Table 2. The positive-definite matrix data example: the leave-one-out cross-validation prediction error under two SDR methods and three metrics.

	Affine invariant	S-divergence	Euclidean
MSIR	0.306	0.302	0.333
GSIR	0.319	0.328	0.357

where $|\cdot|$ denotes the determinant; the symmetrized Kullback–Leibler divergence, $d_{KL}(M_1, M_2) = \{h(M_1, M_2) + h(M_2, M_1)\}/2$, where $h(M_1, M_2) = \{\mathrm{tr}(M_1^{-1}M_2) + \log|M_1| - \log|M_2|\}/2$; the standard Euclidean metric, $d_E(M_1, M_2) = \|M_1 - M_2\|_F$; and the Pearson metric, $d_P(M_1, M_2) = \|M_1/\|M_1\|_F - M_2/\|M_2\|_F\|_F$. The first three distance metrics properly acknowledge the geometry of the matrix space $\mathcal{M}_d$, the fourth hinges on the normality distribution, and the last two leverage only Euclidean geometry.

Figure 2 shows the first two estimated sufficient predictors graphically. The first sufficient predictors found by MSIR and GSIR are both able to separate the two groups of subjects to a good extent, with MSIR achieving better separation, in general, than that of GSIR. Moreover, the first three distance metrics achieve better separation than the final three metrics, which agrees with our expectation. Table 2 reports the leave-one-out cross-validation prediction error when applying a quadratic discriminant analysis classifier to the first two sufficient predictors. For simplicity, we consider only three metrics: the affine invariant metric and S-divergence metric, owing to their competitive performance, as shown in Figure 2, and the Euclidean metric, which serves as a benchmark. The results confirm the visual observation from Figure 2 that MSIR outperforms GSIR, and that the metrics that acknowledge the matrix geometry outperform those that do not.

6.3. Compositional data

As the final example, we consider a compositional data set from a gut microbiota study (Guo et al. (2016)). The data set consists of $n = 83$ subjects, among whom 41 suffer from gout, and the rest do not. For each subject, $p = 3{,}684$ operational taxonomic units (OTUs) were measured, which together characterize the structure of the subject's intestinal microbiota. It is of scientific interest to understand the association between the gout status and the OTU compositions (Guo et al. (2016)). Thus, we aim to produce sufficient predictors that correctly reflect the gout status of a subject.

We follow the data-processing procedure of Pan et al. (2020), who analyzed the same data. Specifically, we first standardize the OTUs, so that the OTU

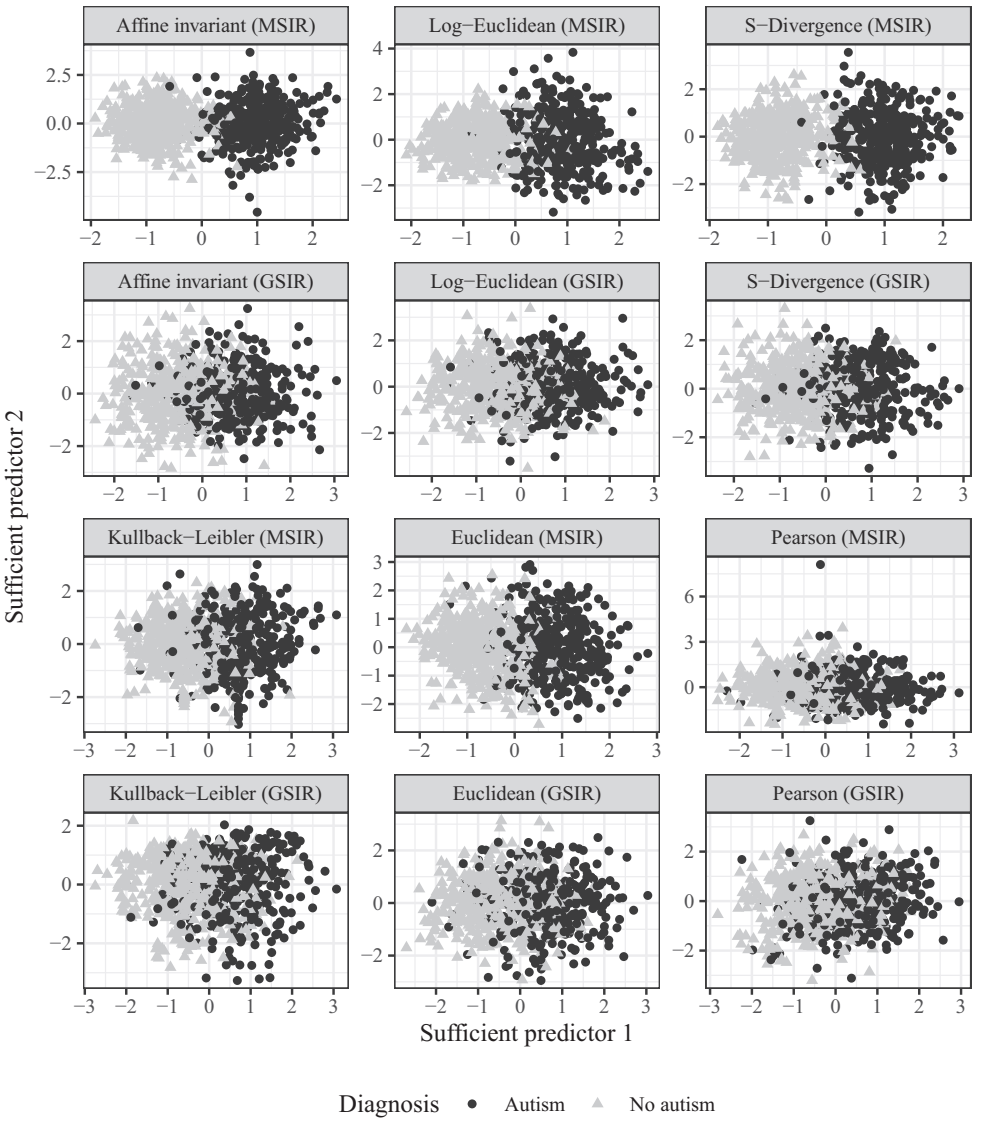

Figure 2. The positive-definite matrix data example: the sufficient predictors under two SDR methods and six metrics, with two groups of subjects, namely, autism and control, marked by different colors.

measurements for each subject sum to one, and thus the data are compositional. In addition, the data are highly sparse, in that, on average, only 202 out of 3,684 measurements are nonzero. As in Pan et al. (2020), we map the standardized vector to the p-dimensional unit sphere by taking element-wise square roots of the coordinates.

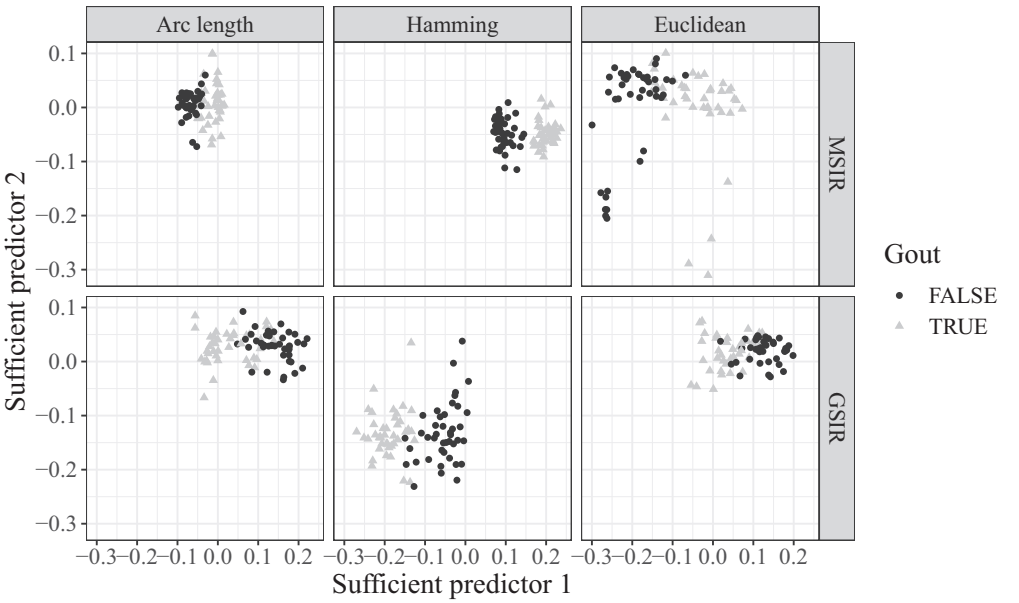

Figure 3. The compositional data example: the sufficient predictors under two SDR methods (row) and three metrics (column), with two groups of subjects, those with gout or not, marked by different colors.

We consider three distance metrics. The first metric is the arc length distance between two transformed compositions. The second metric is the Hamming distance, evaluated on the dichotomized transformation of the compositions; that is, the nonzero entries all become equal to one. This is motivated by the observation that the compositions are very sparse, and that the positions rather than the magnitudes of the nonzero entries are more relevant. The third metric is the usual Euclidean distance.

Figure 3 shows the estimated top two sufficient predictors graphically. Here, the first sufficient predictors found by MSIR and GSIR are both able to separate the two groups of subjects, to some extent. MSIR with the Hamming distance metric achieves the best separation. Table 3 reports the leave-one-out cross-validation prediction error when applying a quadratic discriminant analysis classifier to the extracted sufficient predictors when d is taken as one and then two. Again, the proposed MSIR with the Hamming distance metric achieves the best prediction accuracy. Moreover, there is little difference between $d = 1$ and $d = 2$, suggesting that a single summary predictor is sufficient, which agrees with our expectation, because the response is only binary.

Table 3. The compositional data example: the leave-one-out cross-validation prediction error under two SDR methods, three metrics, and two working dimensions.

d	Method	Arc length	Hamming	Euclidean
1	MSIR	0.241	0.229	0.253
	GSIR	0.253	0.229	0.289
2	MSIR	0.229	0.229	0.277
	GSIR	0.253	0.229	0.289

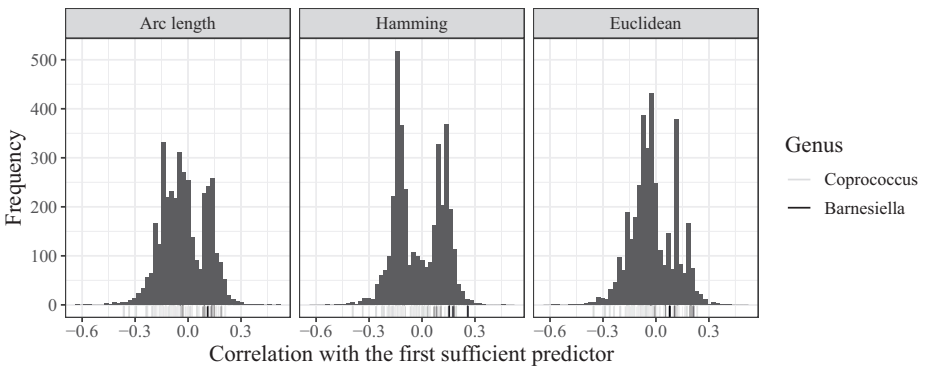

Figure 4. The compositional data example: histograms of the correlations between the first sufficient predictor obtained using MSIR and the original predictor under the three metrics.

To conclude this paper, we give an example on how to interpret the obtained sufficient predictors. The key idea is to compute the correlations between the sufficient and the original predictors. Figure 4 shows histograms of the correlations between the first sufficient predictor obtained using MSIR and the original predictor under the three metrics, which demonstrate a relatively clear bimodal pattern. By Figure 3, a large value of the first MSIR sufficient predictor indicates the presence of gout in a subject. As such, we expect the rightmost peaks of the three histograms in Figure 4 to correspond to OTUs associated with gout. To confirm this, we note that Guo et al. (2016) identified the OTUs of the geni *Coprococcus* (78 in total) and *Barnesiella* (14 in total) as those most associated with subjects without gout and those with gout, respectively. The OTUs of these two geni have been colored in the rugs below the histograms of Figure 4, and are indeed roughly divided between the two modes of the histograms, with *Coprococcus* concentrating to the left peak and *Barnesiella* to the right. This effect is most pronounced in the middle histogram, corresponding to the Hamming distance, which is in line with our result that the Hamming distance gives the best performance of the three distance metrics.

Supplementary Material

The Supplementary Appendix contains the proofs of our theoretical results.

Acknowledgments

The authors thank the editor, associate editor, and anonymous referee for their constructive comments. Virta's research was supported by the Academy of Finland (Grant 335077). Lee's research was partially supported by the NSF grant CIF-2102243 and the Seed Funding grant from the Fox School of Business, Temple University. Li's research was partially supported by the NSF grant CIF-2102227 and the NIH grant R01AG061303.

References

Cook, R. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics* **30**, 455–474.

Cook, R. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 328–332.

Cornea, E., Zhu, H., Kim, P., Ibrahim, J. G. and the Alzheimer's Disease Neuroimaging Initiative (2017). Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 463–482.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K. et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry* **19**, 659–667.

Douglas, R. G. (1966). On majorization, factorization, and range inclusion of operators on Hilbert space. In *Proceedings of the American Mathematical Society* **17**, 413–415.

Dubey, P. and Müller, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106**, 803–821.

Fox, M. D. and Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Frontiers in Systems Neuroscience* **4**, 1–13.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research* **5**, 73–99.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics* **37**, 1871–1905.

Guo, Z., Zhang, J., Wang, Z., Ang, K. Y., Huang, S., Hou, Q. et al. (2016). Intestinal microbiota distinguish gout patients from healthy humans. *Scientific Reports* **6**, 1–10.

Hein, M. and Bousquet, O. (2004). Kernels, associated structures and generalizations. Technical report. Max Planck Institute for Biological Cybernetics.

Hung, H. and Huang, S.-Y. (2019). Sufficient dimension reduction via random-partitions for the large-p-small-n problem. *Biometrics* **75**, 245–255.

Lee, K.-Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics* **41**, 221–249.

Lee, K.-Y. and Li, L. (2022). Functional sufficient dimension reduction through average Fréchet

derivatives. *The Annals of Statistics* **50**, 904–929

Li, B. (2018a). Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics* **46**, 79–103.

Li, B. (2018b). *Sufficient Dimension Reduction: Methods and Applications with R*. Chapman and Hall, CRC.

Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics* **36**, 3182–3210.

Li, B. and Song, J. (2017). Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics* **45**, 1059–1095.

Li, B. and Song, J. (2022). Dimension reduction for functional data based on weak conditional moments. *The Annals of Statistics* **50**, 107–128.

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

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

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association* **87**, 1025–1039.

Lin, L., Thomas, B. S., Zhu, H. and Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* **112**, 1261–1273.

Lin, Z. and Yao, F. (2019). Intrinsic Riemannian functional data analysis. *The Annals of Statistics* **47**, 3533–3577.

Lu, J., Shi, P. and Li, H. (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **75**, 235–244.

Luo, R., Wang, H., Tsai, C.-L. et al. (2009). Contour projected dimension reduction. *The Annals of Statistics* **37**, 3743–3778.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107**, 168–179.

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics* **41**, 250–268.

Pan, W., Wang, X., Zhang, H., Zhu, H. and Zhu, J. (2020). Ball covariance: A generic measure of dependence in Banach space. *Journal of the American Statistical Association* **115**, 307–317.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics* **47**, 691 – 719.

Rudie, J., Brown, J., Beck-Pancer, D., Hernandez, L., Dennis, E., Thompson, P. et al. (2013). Altered functional and structural brain network organization in autism. *NeuroImage: Clinical* **2**, 79 – 94.

Sra, S. (2016). Positive definite matrices and the S-divergence. In *Proceedings of the American Mathematical Society* **144**, 2787–2797.

Sun, W. and Li, L. (2017). Sparse tensor response regression and neuroimaging analysis. *Journal of Machine Learning Research* **18**, 4908–4944.

Tomassi, D., Forzani, L., Duarte, S. and Pfeiffer, R. M. (2019). Sufficient dimension reduction for compositional data. *Biostatistics* **22**, 687–705.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N.

et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–289.

Wang, H. and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics* **35**, 1849–1873.

Xia, Q., Xu, W. and Zhu, L. (2015). Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica* **25**, 1025–1044.

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

Yeh, Y.-R., Huang, S.-Y. and Lee, Y.-J. (2008). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering* **21**, 1590–1603.

Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis* **99**, 1733–1757.

Ying, C. and Yu, Z. (2020). Fréchet sufficient dimension reduction for random objects. *arXiv preprint arXiv:2007.00292*, 1–64.

Zhang, J., Sun, W. W. and Li, L. (2020). Mixed-effect time-varying network model and application in brain connectivity analysis. *Journal of the American Statistical Association* **115**, 2022–2036.

Zhu, H., Chen, Y., Ibrahim, J. G., Li, Y., Hall, C. and Lin, W. (2009). Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association* **104**, 1203–1212.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–643.

Joni Virta

Department of Mathematics and Statistics, University of Turku, FI20014 Turun yliopisto, Finland.

E-mail: jomivi@utu.fi

Kuang-Yao Lee

Department of Statistical Science, Temple University, Philadelphia, PA 19122, USA.

E-mail: kuang-yao.lee@temple.edu

Lexin Li

School of Public Health, University of California at Berkeley, Berkeley, CA 94720, USA.

E-mail: lexinli@berkeley.edu

(Received March 2022; accepted May 2022)