AUTHOR	Joni Virta, Kuang-Yao Lee and Lexin Li
TITLE	Sliced Inverse Regression in Metric Spaces
YEAR	2022
DOI	10.5705/ss.202022.0097
VERSION	Author's accepted manuscript
CITATION	Joni Virta, Kuang-Yao Lee and Lexin Li 2022, Sliced Inverse Regression in Metric Spaces. Statistica Sinica 32 (2022), 2315-2337. DOI 10.5705/ss.202022.0097

Statistica Sinica

SLICED INVERSE REGRESSION IN METRIC SPACES

Joni Virta[1], Kuang-Yao Lee[2], and Lexin Li[3]

[1] *University of Turku*

[2] *Temple University*

[3] *University of California at Berkeley*

Abstract: In this article, we propose a general nonlinear sufficient dimension reduction (SDR) framework when both the predictor and response lie in some general metric spaces. We construct reproducing kernel Hilbert spaces whose kernels are fully determined by the distance functions of the metric spaces, then leverage the inherent structures of these spaces to define a nonlinear SDR framework. We adapt the classical sliced inverse regression of Li (1991) within this framework for the metric space data. We build the estimator based on the corresponding linear operators, and show it recovers the regression information unbiasedly. We derive the estimator at both the operator level and under a coordinate system, and also establish its convergence rate. We illustrate the proposed method with both synthetic and real datasets exhibiting 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 prevailing in almost every branch of science and business, whereas dimension reduction plays a central role in the analysis of such data. A particularly useful reduction paradigm is *sufficient dimension reduction* (SDR), which embodies a family of methods that aim to reduce the dimensionality while losing no information in a regression setting. Since the pioneering work of *sliced inverse regression* (Li, 1991, SIR), SDR has enjoyed a rapid development in the past three decades. 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 since d is often much smaller than p, dimension reduction is achieved. SDR then seeks the minimum subspace spanned by β, called the *central subspace*, which uniquely exists under very mild conditions (Yin et al., 2008). Originating from SIR (Li, 1991), there has been a large body of methods proposed for SDR, mostly in a model-free fashion that does not impose any specific parametric form for the association between Y and $\beta^\intercal X$. Examples include Cook and Weisberg (1991); Li (1992); Cook and Li (2002); Xia et al. (2002); Li and Wang (2007); Ma and Zhu (2012, 2013), among many others.

See also Li (2018b) for a comprehensive review.

SDR in (1.1) achieves *linear dimension reduction*, as the low-dimensional representation takes the form of linear combinations of X. It preserves the original coordinates of X and is easier to interpret; nevertheless, it is less flexible. A more recent line of SDR research instead seeks *nonlinear dimension reduction* (Fukumizu et al., 2004, 2009; Li et al., 2011; Lee et al., 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 a smaller number of functions than its linear counterpart to capture the full regression information, though it is generally harder to interpret.

Despite the substantial progress of SDR, most existing SDR solutions target data that reside 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 a wide range of applications, such as medical imaging, computational biology, and computer vision, and it is of great interest to understand associations among those 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 in this article. For instance, in applications such as brain structural and functional connectivity analysis (Zhu et al., 2009; Zhang et al., 2020), the data usually come in

the form of positive definite matrices, which measure the connectivity strengths of pairs of nodes of a network and admit a certain manifold structure. In applications such as chemistry, geology, and microbiome analysis (Lu et al., 2019), the data are the proportions of individual components that sum to a fixed constant. Meanwhile, 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 in each case to characterize the intrinsic features of the data.

In this article, we propose a general nonlinear SDR framework when both the predictor and response lie in some general, and possibly different, metric spaces. Our key idea is to construct a pair of reproducing kernel Hilbert spaces (RKHS), whose kernels 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, and further adapt sliced inverse regression of Li (1991) within this framework. We build the estimator based on some linear operators, and show it recovers the regression information unbiasedly. 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 with both synthetic and real datasets exhibiting non-Euclidean geometry.

Our proposal is related to but also clearly differs from the nonlinear SDR method

of Lee et al. (2013), and some recent SDR solutions involving functional or non-Euclidean data such as Yeh et al. (2008); Li and Song (2017); Tomassi et al. (2019); Ying and Yu (2020); Lee and Li (2022). In particular, Lee et al. (2013) developed a general framework for nonlinear SDR and proposed to estimate the functions $f_1, \ldots, f_d$ in (1.2) as the eigenfunctions of some linear operator defined on a Hilbert space $\mathcal{H}$, but they still targeted the Euclidean data. Besides, they took $\mathcal{H}$ 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 aim at data residing in a general metric space. Moreover, we take $\mathcal{H}$ to be an RKHS at both the population and sample levels, which makes the connection between the population and sample versions of the estimation procedure more transparent. Yeh et al. (2008) proposed kernel SIR under the framework of (1.2), but required 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. Relatedly, Lee and Li (2022) studied linear SDR when both X and Y are functions in some Hilbert space. By contrast, we consider more general data objects than functional data. Tomassi et al. (2019) developed linear SDR for compositional data, but restricted 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. Since the dimension reduction is to be performed for the predictors, our method differs considerably from that of Ying and Yu (2020).

The rest of the article is organized as follows. Section 2 develops the general framework of nonlinear SDR for data in metric spaces, and Section 3 derives the metric version of SIR under this framework. Section 4 describes the finite-sample implementation, and Section 5 studies the convergence properties of the estimator. Section 6 presents the numerical studies, and the Supplementary Appendix collects all the technical proofs.

2. Nonlinear SDR for Metric Space Data

In this section, we propose a general framework for conducting nonlinear SDR for data residing in arbitrary metric spaces. It involves three main steps: defining a minimal σ-field that captures full regression information, constructing reproducing kernel Hilbert spaces of both X and Y from the metric spaces, and using the 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), (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, Ω_Y^0, there may be multiple feasible choices for the metrics d_X, d_Y. See, for instance, Section 6 where 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 topology in Ω_X^0 and Ω_Y^0, respectively. Consider $X : \Omega \to \Omega_X^0$ that is a $\mathcal{F}/\mathcal{F}_X$-measurable random variable with the distribution $P_X = P \circ X^{-1}$, and $Y : \Omega \to \Omega_Y^0$ that is a $\mathcal{F}/\mathcal{F}_Y$-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 et al. (2013).

Definition 1. A sub-σ-field $\mathcal{G}$ of σ_X is said to be a sufficient dimension reduction 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. It is called the central σ-field, and denoted by $\mathcal{G}_{Y|X}$.

Definition 1 suggests that there exists uniquely a 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 dimension reduction is achieved 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 $f : \Omega_X^0 \to \mathbb{R}$ and, indeed, this is what Lee et al. (2013) used.

We instead take $\mathcal{H}_X$ to be a suitably defined reproducing kernel Hilbert space, a choice that makes the subsequent methodology and theory development considerably simpler. More specifically, 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 $x_1, x_2 \in \Omega_X^0$,

$$\kappa(x_1, x_2) = \rho\{d_X(x_1, x_2)\}, \tag{2.3}$$

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, for instance, the Gaussian kernel and the Laplace kernel, among others. Throughout our implementation, we employ the 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,

$$\langle f, \mu_X \rangle_{\mathcal{H}_X^0} = \mathrm{E}\{f(X)\}, \quad \text{for all} \quad f \in \mathcal{H}_X^0,$$

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

Note that every $f_0 \in \ker(\Sigma^0_{XX})$ satisfies that $\mathrm{Var}\{f_0(X)\} = \langle f_0, \Sigma^0_{XX} f_0 \rangle_{\mathcal{H}^0_X} = 0$, and is almost surely equal to a constant, where $\ker(\cdot)$ denotes the null space. As such, we further restrict our attention to $\mathcal{H}_X = \overline{\mathrm{ran}}(\Sigma^0_{XX})$, 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 $\Omega_X \subseteq \Omega^0_X$, 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 reveals 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 through the inner product, $\langle f, \kappa_X(\cdot, x) - \mu_X \rangle_{\mathcal{H}_X} = f(x) - \mathrm{E}\{f(X)\}$. Its proof also shows that the space $\mathcal{H}_X$ admits an alternative characterization, i.e., $\mathcal{H}_X = \overline{\mathrm{span}}\{\kappa_X(\cdot, x) - \mu_X : x \in \Omega_X\}$, where $\overline{\mathrm{span}}(\cdot)$ denotes the closure of the say spanned by the set of functions. We briefly remark that a similar result was obtained in Li and Song (2017, Lemma 1). However, their proof implicitly assumed that the only set for which P_X assigns a zero probability is the empty set, essentially ruling out all continuous distributions, whereas our Lemma 1 fixes this issue. We also remark that, 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 in our dimension reduction pursuit. 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, i.e., the central σ-field. In our setting, the concept of "dimensionality" is less obvious than that in the classical SDR setting, which is simply the dimension of the central subspace. This is because there are sets that generate the same σ-field, but with very different dimensions. Nevertheless, our formulation is useful when one is interested in reducing the dimensionality in the class sense, as the central class induced by the central σ-field contains all such sets of functions generating 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 Sliced Inverse Regression

In this section, we derive the population-level sliced inverse regression for metric space data. Recall the classical SIR (Li, 1991) when both X and Y lie in an Euclidean space. It estimates the central subspace by the range of the matrix,

$$\mathrm{Var}(X)^{-1}\mathrm{Var}\{\mathrm{E}(X \mid Y)\}, \tag{3.4}$$

We next derive the operator analogue for (3.4) 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 main building blocks of our nonlinear metric SIR procedure.

$$\begin{aligned}
&\Sigma_{XX} : \mathcal{H}_X \to \mathcal{H}_X, \quad \langle f, \Sigma_{XX} f' \rangle_{\mathcal{H}_X} = \mathrm{Cov}\{f(X), f'(X)\}, \\
&\Sigma_{XY} : \mathcal{H}_Y \to \mathcal{H}_X, \quad \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_X} = \mathrm{Cov}\{f(X), g(Y)\}, \\
&\Sigma_{YY} : \mathcal{H}_Y \to \mathcal{H}_Y, \quad \langle g', \Sigma_{YY} g \rangle_{\mathcal{H}_Y} = \mathrm{Cov}\{g'(Y), g(Y)\},
\end{aligned} \tag{3.5}$$

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}. We also note that, 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 generally holds, e.g., when κ_X is a Gaussian kernel (Fukumizu et al., 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), and similar conditions are

commonly imposed in SDR (Ying and Yu, 2020; Li and Song, 2022). It guarantees that the operator $\Sigma_{YY}^{\dagger}\Sigma_{YX}$ is both well-defined and bounded (Douglas, 1966, Theorem 1), where $\dagger$ denotes the Moore-Penrose pseudo-inverse of Σ_{YY}; see Li (2018a) for more 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. They are essential to construct the operator analogue for the SIR estimator (3.4). In addition, they help turn conditional moments into unconditional ones, avoiding the slicing step in the original SIR.

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

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

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

By Lemma 2, the operator $\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}$ can be seen as the analogue of the matrix $\mathrm{Var}\{\mathrm{E}(X \mid Y)\}$ in (3.4). Besides, the operator $\Sigma_{XX}^{\dagger}$ can be seen as the analogue of $\mathrm{Var}(X)^{-1}$ in (3.4). Consequently, a direct operator counterpart of (3.4) is,

$$\Lambda_{\mathrm{SIR}} = \Sigma_{XX}^{\dagger}\Sigma_{XY}\Sigma_{YY}^{\dagger}\Sigma_{YX}. \tag{3.6}$$

This operator is well-defined by Assumption 3. Moreover, it is interesting to note that, if we choose linear kernels κ_X, 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, which is parallel to the classical SIR for linear SDR of Euclidean data. We need an additional regularity condition.

Assumption 4. Suppose the set $\text{ran}(\Sigma_{XX}) \cap \mathcal{S}_{Y|X}^{\perp}$ 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 $\text{ran}(\Sigma_{XX})$ and $\mathcal{S}_{Y|X}^{\perp}$ is suitably rich in $\mathcal{S}_{Y|X}^{\perp}$, which is a mild condition, since $\text{ran}(\Sigma_{XX})$ is, by definition, dense in its closure $\mathcal{H}_X$. Similar condition has been imposed implicitly in Li and Song (2017).

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

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

3.2 Discrete response

Next, we consider a special case when Y lies in the usual Euclidean space and is discrete. This is the scenario that 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, instead of resorting to the unconditional representations as in Lemma 2.

More specifically, suppose $\Omega_Y^0 = \{1, \ldots, K\}$, and let $\pi_k = P(Y = k)$, $\pi_k > 0$ for all $k \in \Omega_Y^0$. By the Riesz representation theorem, there exists the elements $\gamma_{X|k} \in \mathcal{H}_X$, $k = 1, \ldots, K$, 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 $\gamma_{X|k}$ can be seen to provide a discrete counterpart of Lemma 2(a). We then define the covariance operator,

$$\Gamma_{XX|Y} = \sum_{k=1}^{K} \pi_k (\gamma_{X|k} \otimes \gamma_{X|k}) : \mathcal{H}_X \to \mathcal{H}_X, \tag{3.7}$$

where $\otimes$ denotes the tensor product. It satisfies 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 (3.6) when Y is categorical is,

$$\Lambda_{\mathrm{SIR,D}} = \Sigma_{XX}^{\dagger} \Gamma_{XX|Y}. \tag{3.8}$$

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, 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 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 by $\hat{\mu}_X = \mathrm{E}_n\{\kappa_X(\cdot, X)\}$, 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 $\Sigma_{XX}, \Sigma_{XY}, \Sigma_{YY}$ in (3.5) as,

$$\hat{\Sigma}_{XX} = \mathrm{E}_n[\{\kappa_X(\cdot, X) - \hat{\mu}_X\} \otimes \{\kappa_X(\cdot, X) - \hat{\mu}_X\}],$$

$$\hat{\Sigma}_{XY} = \mathrm{E}_n[\{\kappa_X(\cdot, X) - \hat{\mu}_X\} \otimes \{\kappa_Y(\cdot, Y) - \hat{\mu}_Y\}],$$

$$\hat{\Sigma}_{YY} = \mathrm{E}_n[\{\kappa_Y(\cdot, Y) - \hat{\mu}_Y\} \otimes \{\kappa_Y(\cdot, Y) - \hat{\mu}_Y\}].$$

Moreover, we have $\hat{\Sigma}_{YX} = \hat{\Sigma}^*_{XY}$. We then obtain the sample estimator of the metric SIR operator Λ_{SIR} in (3.6) 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 utilize the ridge regularization to estimate the pseudo-inverses, τ_1, τ_2 are the ridge parameters, and I is the identity operator. Finally, we estimate the range of

Λ_{SIR} through 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 (3.7) 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 (3.8) as,

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

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

4.2 Estimation under a coordinate representation

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

In Step 1, we choose the kernel function κ_X and κ_Y. There are multiple choices of kernel functions, while we employ the Gaussian kernel throughout our implementation. We use the leave-one-out cross-validation to determine the bandwidth parame-

ters in κ_X and κ_Y, following a similar strategy as in Lee et al. (2013). We then compute the Gram matrix $K_X = (\kappa_X(X_i, X_{i'}))_{i,i'=1}^n \in \mathbb{R}^{n\times n}$, 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, d_Y as in (2.3). 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

$$G_X = QK_XQ, \quad \text{and} \quad G_Y = QK_YQ. \tag{4.9}$$

In Step 2, we compute the coordinate representation of the sample metric SIR operator $\hat{\Lambda}_{\text{SIR}}$. Toward that end, consider the sample counterpart of the space $\mathcal{H}_X^0$, which is the span of the sample elements, $\hat{\mathcal{H}}_X^0 = \text{span}\big\{\kappa_X(\cdot, X_i) \mid i = 1, \ldots, n\big\}$. 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 also construct a linearly independent basis via Karhunen-Loève expansion, see, e.g. Lee and Li (2022).

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

Under Assumption 6, the elements $\kappa_X(\cdot, X_i)$, $i = 1, \ldots, n$, 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, $\hat{\mathcal{H}}_X = \text{span}\big\{\kappa_X(\cdot, X_i) - \hat{\mu}_X \mid i = 1, \ldots, n\big\}$. We construct the sample spaces $\hat{\mathcal{H}}_Y^0$ and $\hat{\mathcal{H}}_Y$ similarly.

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

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

where G_X, G_Y are as defined in (4.9). We also clarify that, the above coordinate representation seems to suggest that $\hat{\Sigma}_{YX}$ does not depend on Y, which is not the case. Actually, $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YX}$ share the same coordinate, which is $n^{-1}G_X$, but they involve two different sets of bases, as $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YX}$ have different range spaces. For simplicity, we drop the involvement of the underlying bases in the coordinate bracket notation. But we remind that $\hat{\Sigma}_{YX}$ depends on Y through the underlying bases. A similar discussion applies to $\hat{\Sigma}_{XY}$ too.

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

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

$$[\hat{\Lambda}_{\text{SIR}}] = G_X^\dagger G_Y G_Y^\dagger G_X, \tag{4.10}$$

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

To improve 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 $G_Y^\dagger$ by $\{G_Y + \tau_2 I_n\}^{-1}$ with $\tau_2 = c \times \phi_1(G_Y)$.

In Step 3, we estimate the range of $\hat{\Lambda}_{\text{SIR}}$ through the eigen-decomposition of its coordinate in (4.10). Letting $v_1, \ldots, v_d$ denote the d leading eigenvectors of $[\hat{\Lambda}_{\text{SIR}}][\hat{\Lambda}_{\text{SIR}}]^\intercal$, 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 $[\hat{\Lambda}_{\text{SIR}}]$.

We remark that, the computational complexity of our proposed method is of the order $\mathcal{O}(n^3)$. 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 the 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 for possible discrepancy in the signs of the resulting eigenvectors, we choose their signs such that, for each $j = 1, \ldots, d$, the sum $\sum_{q,q'=1}^{Q} v_{j,q}^\intercal v_{j,q'}$ is maximized, where $v_{j,q}$ is the jth eigenvector of $[\hat{\Lambda}_{\text{SIR}}][\hat{\Lambda}_{\text{SIR}}]^\intercal$ 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 samples.

For the special case when 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 $\mathbb{I}(Y_i = k)$, $i = 1, \ldots, n$. 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 SDR. There have been a number of information criterion-based selection proposals for SDR of the Euclidean data (Zhu et al., 2006; Luo et al., 2009; Xia et al., 2015). We expect a similar information criterion is applicable for our metric SIR as well, while we leave the full investigation as future research.

5. Asymptotic Theory

In this section, we establish the convergence rate of the proposed metric SIR estimator at the operator level for both 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, Ω_Y^0, it ensures that the RKHS $\mathcal{H}_X, \mathcal{H}_Y$ are separable (Hein and Bousquet, 2004), which in turn ensures that $\mathcal{H}_X, \mathcal{H}_Y$ 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 seen as a stronger version of Assumption 3; that is, in comparison with Assumption 3, the mapping of Σ_{XY} needs to concentrate even more on the leading eigen-spaces of Σ_{XX} and Σ_{YY}. This, again, can be understood as a smoothness condition.

In our sample estimation, we employ the ridge regularization for the pseudo-inverses. For simplicity, in our theoretical analysis, we suppose the ridge parameters $\tau_1 = \tau_2 = \tau$, and 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 $\beta = 1/4$. We remark 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 study the empirical performance of our proposed metric sliced inverse regression (MSIR), under difference choices of distance metrics. We also compare with the nonlinear SIR method of Lee et al. (2013, GSIR). Although GSIR was originally formulated through the Euclidean geometry, it can be easily extended to incorporate an arbitrary distance metric.

6.1 Torus manifold data

As the first example, we consider a two-dimensional torus as the predictor, while we simulate the response using different distance metrics. A torus is best visualized as

a unit square $[0,1]^2$ for which the opposite edges have been "glued together". We consider two different generative models.

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

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

where the 2-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. Since the point $(0.5, 0.5)^\intercal$ is 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. The same is not true for Model 2, however, where the reference point $(1,1)^\intercal$ lies at the corner of the square. This means that the true regression relationship is not a smooth function of the Euclidean distance, but it 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, the geodesic distance and the Euclidean distance.

Table 1: The torus data example: the average distance correlation (with the standard deviation in the parenthesis) between the response and 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)

Table 1 reports the distance correlation between the response and the first two estimated sufficient predictors evaluated on the testing samples and averaged over 200 data replications. It is seen that the proposed MSIR outperforms the competing GSIR, by achieving a higher distance correlation and a smaller standard error. More-

over, the Euclidean metric is slightly better suited to Model 1, where the toroidal geometry plays no role, while the geodesic metric is considerably better for Model 2, where the toroidal geometry plays a crucial role. The increased sample size mostly helps to reduce the standard error of the estimator. Figure 1 further provides a visualization of the estimated sufficient predictors for a single data replication under Model 2 with $n = 500$ and $\sigma = 0.05$. It agrees with the qualitative patterns observed in Table 1 that MSIR produces more informative sufficient predictors than 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 dataset consists of $n = 795$ subjects, among whom 362 were diagnosed with autism, and the rest 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 through 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 of both disorder pathology

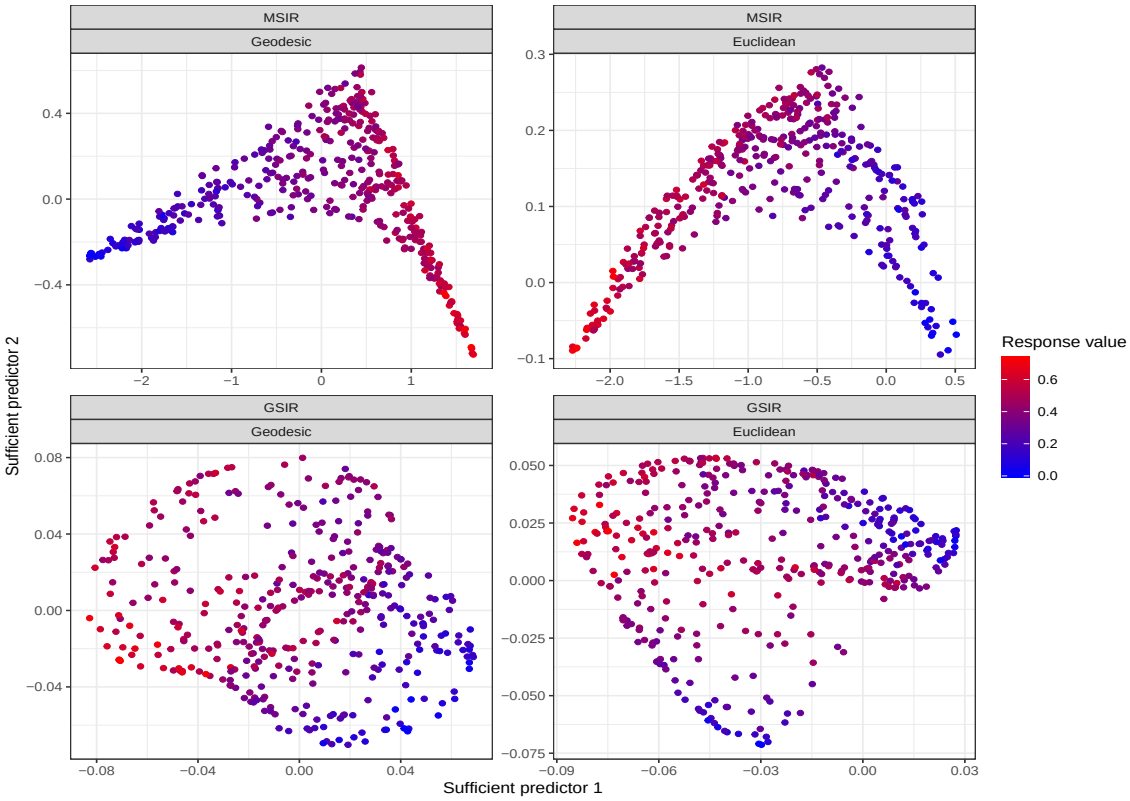

Figure 1: The torus data example: the sufficient predictors under two SDR methods and two distance metrics.

and development of the brain (Fox and Greicius, 2010). It is thus of great scientific importance to understand the association between the autism status and the brain connectivity network, and our goal is to produce sufficient predictors to separate the autism patients from those healthy controls.

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 common principal components analysis, and project the connectivity matrices to 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. These include 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 $\|\cdot\|_F$ 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), $d_S(M_1, M_2) = \log|(M_1+M_2)/2| - (1/2)\log|M_1 M_2|$, 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, $d_P(M_1, M_2) = \|M_1/\|M_1\|_F - M_2/\|M_2\|_F\|_F$. Among these six distance metrics, the first three properly acknowledge the geometry of the matrix space $\mathcal{M}_d$, the fourth one hinges on the normality distribution, and the last two only leverage the Euclidean geometry.

Figure 2 shows the first two estimated sufficient predictors graphically. It is seen that the first sufficient predictors found by MSIR and GSIR are both able to sepa-

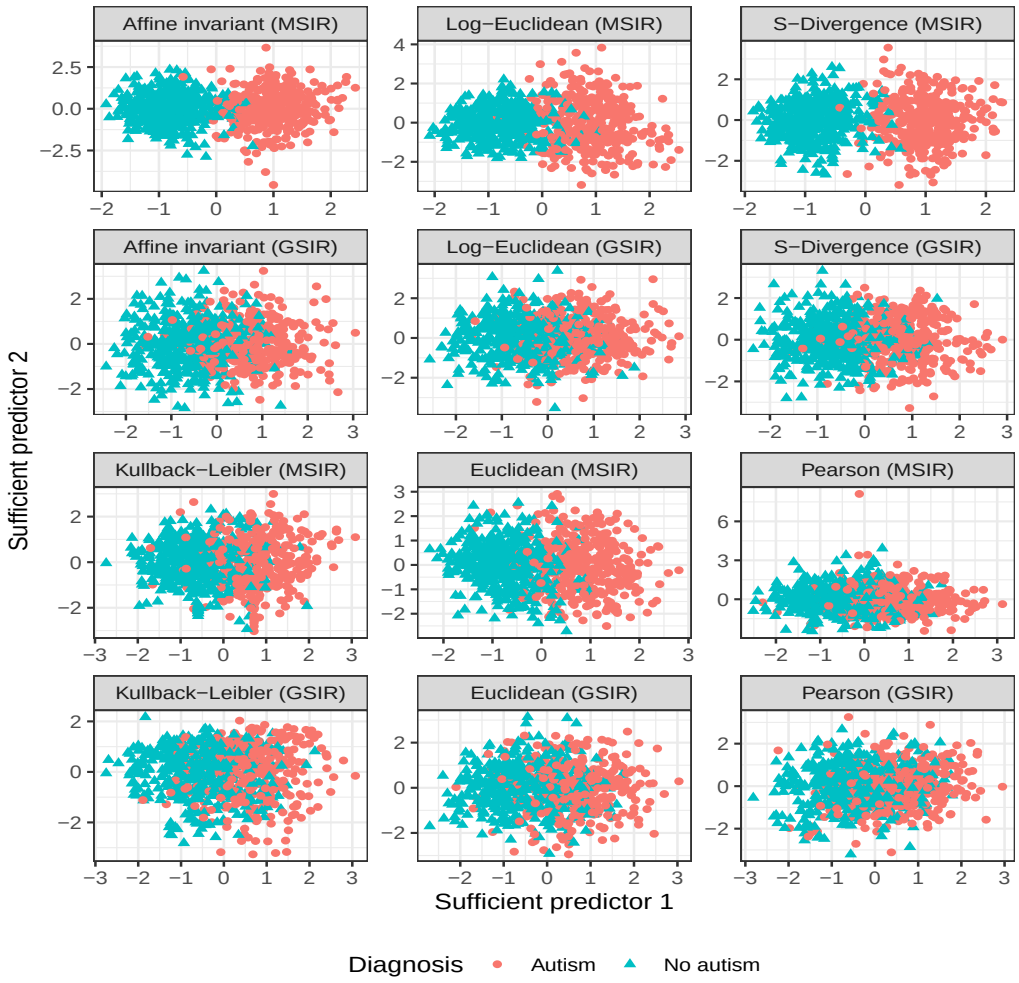

Figure 2: The positive definite matrix data example: the sufficient predictors under two SDR methods and six metrics, with two groups of subjects, autism or control, marked by different colors.

rate the two groups of subjects to a good extent, whereas MSIR achieves generally a better separation than GSIR. Moreover, the first three distance metrics achieve

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

a better separation than the last 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 extracted first two sufficient predictors. For simplicity, we only consider three metrics, the affine invariant metric, and the S-divergence metric, due to their competitive performance as shown in Figure 2, and the Euclidean metric, which serves as a benchmark. It confirms with the visual observation from Figure 2 that MSIR outperforms GSIR, and the metrics that acknowledge the matrix geometry outperform the one that does not.

6.3 Compositional data

As the final example, we consider a compositional dataset from a gut microbiota study (Guo et al., 2016). The dataset consists of $n = 83$ subjects, among whom 41 suffered from gout, and the rest not. For each subject, $p = 3684$ operational taxonomic units (OTUs) were measured, which characterizes the structure of the subject's intestinal

microbiota. It is of great scientific interest to understand the association between the gout status and the OTU compositions (Guo et al., 2016), and we aim to produce sufficient predictors to reflect the gout status.

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 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 3684 measurements are non-zero. 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.

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; i.e., all the nonzero entries are turned into one. This is motivated by the observation that the compositions are very sparse, and 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. It is seen that the first sufficient predictors found by MSIR and GSIR are both able to separate the two groups of subjects to some extent. Particularly, 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

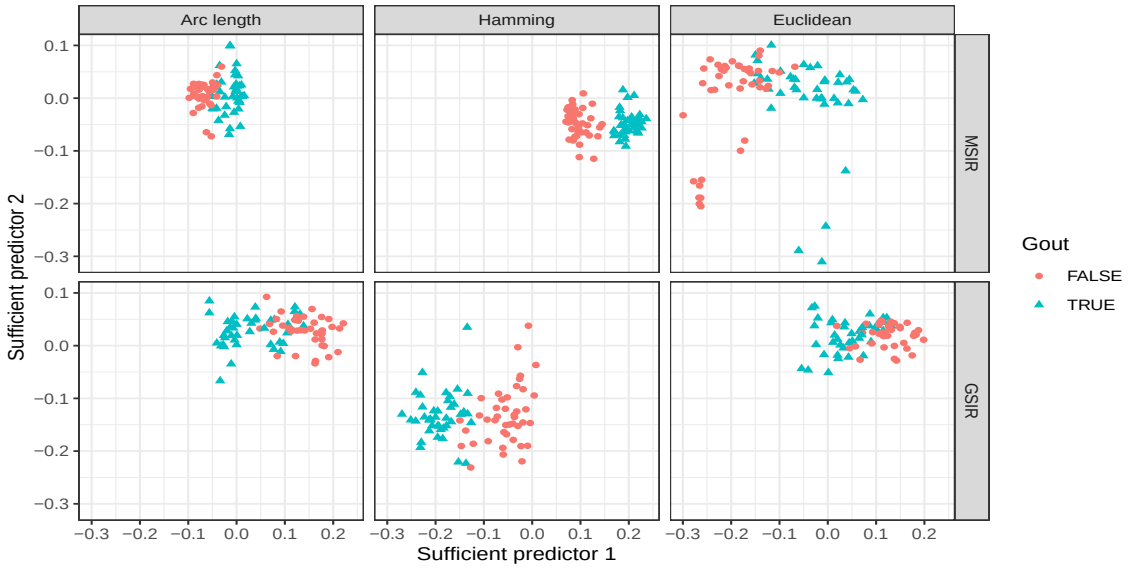

Figure 3: The compositional data example: the sufficient predictors under two SDR methods (row) and three metrics (column), with two groups of subjects, gout or not, marked by different colors.

to the extracted sufficient predictors when d is taken as 1 and 2, respectively. 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 a single summary predictor is sufficient, which agrees with our expectation since the response is only binary.

To conclude this study, we give an example on how to interpret the obtained sufficient predictors. The key idea is to compute the correlations between the sufficient and original predictors. Figure 4 shows the histograms of the correlations

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

between the first sufficient predictor obtained by 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 that are 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 the ones mostly associated with non-gouty and gouty subjects, respectively. The OTUs of these two geni have been colored in the rugs below the histograms of Figure 4 and they 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

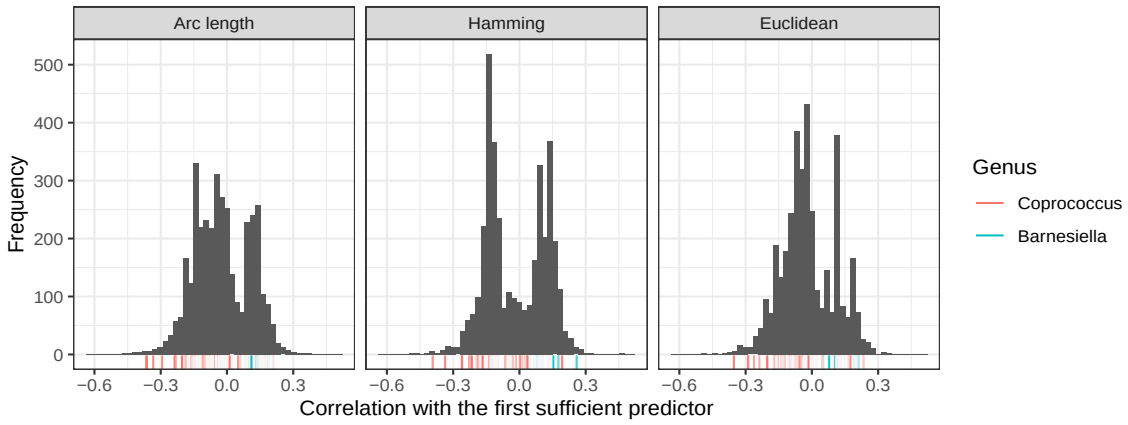

Figure 4: The compositional data example: the histograms of the correlations between the first sufficient predictor obtained by MSIR and the original predictor under the three metrics.

Hamming distance gives the best performance out of the three distance metrics.

Supplementary Materials

The Supplementary Appendix contains the proofs of our theoretical results.

Acknowledgements

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

Li's research is partially supported by the NSF grant CIF-2102227 and the NIH grant R01AG061303.

References

Cook, R. and B. Li (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics 30*(2), 455–474.

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

Cornea, E., H. Zhu, P. Kim, J. G. Ibrahim, 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*(2), 463–482.

Di Martino, A., C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Muller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry 19*(6), 659–667.

Douglas, R. G. (1966). On majorization, factorization, and range inclusion of operators on Hilbert space.

Proceedings of the American Mathematical Society 17(2), 413–415.

Dubey, P. and H.-G. Müller (2019, 10). Fréchet analysis of variance for random objects. *Biometrika 106*(4), 803–821.

Fox, M. D. and M. Greicius (2010). Clinical applications of resting state functional connectivity. *Frontiers in Systems Neuroscience 4*(19), 1–13.

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

Fukumizu, K., F. R. Bach, and M. I. Jordan (2009). Kernel dimension reduction in regression. *The Annals of Statistics 37*(5), 1871–1905.

Guo, Z., J. Zhang, Z. Wang, K. Y. Ang, S. Huang, Q. Hou, X. Su, J. Qiao, Y. Zheng, L. Wang, H. Danliang, J. Xu, Y. K. Lee, and H. Zhang (2016). Intestinal microbiota distinguish gout patients from healthy humans. *Scientific Reports 6*(1), 1–10.

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

Hung, H. and S.-Y. Huang (2019). Sufficient dimension reduction via random-partitions for the large-p-small-n problem. *Biometrics 75*(1), 245–255.

Lee, K.-Y., B. Li, and F. Chiaromonte (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics 41*(1), 221–249.

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

The Annals of Statistics in press, 1–25.

Li, B. (2018a). Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics 46*(1), 79–103.

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

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

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

Li, B. and J. Song (2022). Dimension reduction for functional data based on weak conditional moments. *The Annals of Statistics 50*(1), 107–128.

Li, B. and S. Wang (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*(414), 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*(420), 1025–1039.

Lin, L., B. S. Thomas, H. Zhu, and D. B. Dunson (2017). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association 112*(519), 1261–1273.

Lin, Z. and F. Yao (2019). Intrinsic Riemannian functional data analysis. *The Annals of Statistics 47*(6), 3533–3577.

Lu, J., P. Shi, and H. Li (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics 75*(1), 235–244.

Luo, R., H. Wang, C.-L. Tsai, et al. (2009). Contour projected dimension reduction. *The Annals of Statistics 37*(6B), 3743–3778.

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

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

Pan, W., X. Wang, H. Zhang, H. Zhu, and J. Zhu (2020). Ball covariance: A generic measure of dependence in Banach space. *Journal of the American Statistical Association 115*(529), 307–317.

Petersen, A. and H.-G. Müller (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics 47*(2), 691 – 719.

Rudie, J., J. Brown, D. Beck-Pancer, L. Hernandez, E. Dennis, P. Thompson, S. Bookheimer, and M. Dapretto (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. *Proceedings of the American Mathematical Society 144*(7), 2787–2797.

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

Tomassi, D., L. Forzani, S. Duarte, and R. M. Pfeiffer (2019). Sufficient dimension reduction for compositional data. *Biostatistics 22*(4), 687–705.

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage 15*(1), 273 – 289.

Wang, H. and J. S. Marron (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics 35*(5), 1849 – 1873.

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

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

Yeh, Y.-R., S.-Y. Huang, and Y.-J. Lee (2008). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE transactions on Knowledge and Data Engineering 21*(11), 1590–1603.

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

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

Zhang, J., W. W. Sun, and L. Li (2020). Mixed-effect time-varying network model and application in brain connectivity analysis. *Journal of the American Statistical Association 115*(532), 2022–2036.

Zhu, H., Y. Chen, J. G. Ibrahim, Y. Li, C. Hall, and W. Lin (2009). Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association 104*(487), 1203–1212.

Zhu, L., B. Miao, and H. Peng (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association 101*(474), 630–643.