Statistica Sinica Preprint No: SS-2023-0396	
Title	On the Optimality of Functional Sliced Inverse Regression
Manuscript ID	SS-2023-0396
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0396
Complete List of Authors	Rui Chen, Songtao Tian, Dongming Huang, Qian Lin and Jun S. Liu
Corresponding Authors	Jun S. Liu
E-mails	jliu@stat.harvard.edu
Notice: Accepted author version.	

On the Optimality of Functional Sliced Inverse Regression

Rui Chen[a], Songtao Tian*[b], Dongming Huang[c], Qian Lin[d] and Jun S. Liu†[e]

[a]*Department of Statistics and Data Science, Tsinghua University*
[b]*Department of Mathematical Sciences, Tsinghua University*
[c]*Department of Statistics and Data Science, National University of Singapore*
[d]*Department of Statistics and Data Science, Tsinghua University; Beijing Academy of Artificial Intelligence, Beijing, 100084, China*
[e]*Department of Statistics, Harvard University, 02138*

Abstract: In this paper, we prove that functional sliced inverse regression (FSIR) achieves the optimal (minimax) rate for estimating the central space in functional sufficient dimension reduction problems. First, we provide a concentration inequality for the FSIR estimator of the covariance of the conditional mean. Based on this inequality, we establish the root-n consistency of the FSIR estimator of the image of covariance of the conditional mean. Second, we apply the most widely used truncated scheme to estimate the inverse of the covariance operator and identify the truncation parameter that ensures that FSIR can achieve the optimal minimax convergence rate for estimating the central space. Finally, we conduct simulations to demonstrate the optimal choice of truncation parameter and the estimation efficiency of FSIR. To the best of our knowledge, this is the first paper to rigorously prove the minimax optimality of FSIR in estimating the central space for multiple-index models and general Y (not necessarily discrete).

Keywords and phrases: Central space, Functional data analysis, Functional sliced inverse regression, Multiple-index models, Sufficient dimension reduction.

*Co-first author.
†Corresponding author.

1. Introduction

Sufficient Dimension Reduction (SDR) aims to identify a low-dimensional subspace that captures the most important features of the data that are relevant to the response variable. It is a useful tool for researchers to perform both exploratory data analysis and detailed model developments when the dimension of the covariates is high. Concretely, for a pair of random variables $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$, an *effective dimension reduction* (EDR) subspace is a subspace $\mathcal{S} \subset \mathbb{R}^p$ such that Y is independent of $\boldsymbol{X}$ given $P_{\mathcal{S}}\boldsymbol{X}$ (where $P_{\mathcal{S}}$ denotes the projection operator from $\mathbb{R}^p$ to $\mathcal{S}$), which can be represented as:

$$Y \perp\!\!\!\perp \boldsymbol{X} \mid P_{\mathcal{S}}\boldsymbol{X}. \tag{1.1}$$

SDR targets at estimating the intersection of all EDR subspaces, which is shown to be again an EDR subspace under mild conditions (Cook, 1996). This intersection is often referred to as the *central space* and denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$. To find the central space $\mathcal{S}_{Y|\boldsymbol{X}}$, researchers have developed a variety of methods: *sliced inverse regression* (SIR, Li 1991), *sliced average variance estimation* (SAVE, Cook and Weisberg 1991), *principal hessian directions* (PHD, Li 1992), *minimum average variance estimation* (MAVE, Xia et al. 2009), *directional regression* (DR, Li and Wang 2007), etc. SIR is one of the most popular SDR methods and its asymptotic properties are of particular interest. For more details, readers can refer to Hsing and Carroll (1992); Zhu and Ng (1995); Zhu et al. (2006); Wu and Li (2011); Lin et al. (2018, 2021); Tan et al. (2020); Huang et al. (2023).

There has been a growing interest in statistical modeling of functional data (i.e., the predictors are random functions in some function space $\mathcal{H}$), and researchers have extended existing multivariate SDR algorithms to accommodate this type of data. In a functional SDR algorithm, the space $\mathbb{R}^p$ is replaced by the function space $\mathcal{H}$, a functional EDR subspace can be defined as a subspace $\mathcal{S} \subset \mathcal{H}$ such that (1.1) holds, and the intersection of all functional EDR subspaces can be referred to as the functional central space.

Functional sliced inverse regression (FSIR) proposed by Ferré and Yao (2003) is one of the earliest functional SDR algorithms. They estimate the central space $\mathcal{S}_{Y|\boldsymbol{X}}$ under a *multiple-index model*. The model is mathematically represented as

$$Y = f(\langle \boldsymbol{\beta}_1, \boldsymbol{X} \rangle, \ldots, \langle \boldsymbol{\beta}_d, \boldsymbol{X} \rangle, \varepsilon) \tag{1.2}$$

where $f : \mathbb{R}^{d+1} \to \mathbb{R}$ is an unknown (non-parametric) link function, the predictor $\boldsymbol{X}$ and indices $\boldsymbol{\beta}_i$'s are functions in $L^2[0,1]$, the separable Hilbert space of square-integrable curves on $[0,1]$, and ε is a random noise independent of $\boldsymbol{X}$. Although the individual $\boldsymbol{\beta}_j$'s are unidentifiable because of the flexibility of the link function f, space $\mathcal{S}_{Y|\boldsymbol{X}} := \mathrm{span}\{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d\}$ is estimable. Several key findings concerning FSIR have been established since its introduction. Ferré and Yao (2003) showed the consistency of the FSIR estimator under some technical assumptions. However, they did not establish a similar convergence rate for FSIR as those for the multivariate SIR obtained by Hsing and Carroll (1992) and Zhu and Ng (1995). Forzani and Cook (2007) found that the $\sqrt{n}$-consistency of the central space can not be achieved by the FSIR estimator unless some very restrictive conditions on the covariance operator of the predictor are imposed. Yao et al. (2015) introduced the technique of functional cumulative slicing estimation (FCSE) for SDR, focusing on scenarios with sparse designs, and also obtained its convergence rate. Lian (2015) showed that the convergence rate of the FSIR estimator for a discrete response Y is the same as that for *functional linear regression* in (Hall and Horowitz, 2007), but they did not provide a rigorous proof of the optimality of FSIR, which is deemed quite challenging. Recent developments in this field can be found in Lian and Li (2014); Wang and Lian (2020).

Recently, there have been significant advances in understanding the behavior of SIR for high dimensional data (i.e., $\rho := \lim p/n$ is a constant or ∞). Lin et al. (2018) established the phase transition phenomenon of SIR in high dimensions, i.e., SIR can estimate the central space consistently if and only if $\rho = 0$. Lin et al. (2021) further obtained the minimax convergence rate of estimating the central space in high dimensions and showed that SIR can achieve the minimax rate. In a different setting, Tan et al. (2020) studied the minimax rate in high dimensions under various loss functions and proposed a computationally tractable adaptive estimation scheme for sparse SIR. Huang et al. (2023) generalized the minimax rate results to cases with a large structural dimension d (i.e., there is no constant upper bound on the dimension of the central space d). These work in high dimensional SIR inspires our research to address the theoretical challenges of FSIR and bridge the aforementioned theoretical gap between FSIR and multivariate SIR.

In the present paper, we focus on examining the error bound of FSIR under very mild conditions. We show that FSIR can estimate the central space optimally for general Y (not necessarily discrete) over a large class of distributions.

1.1 Major contributions

Our main results are summarized as follows:

(i) To study the asymptotic properties of FSIR under very general settings, we introduce a fairly mild condition called the *weak sliced stable condition* (WSSC) for functional data (see Definition 1).

(ii) Under the above WSSC, we prove a concentration inequality for the FSIR estimator $\widehat{\Gamma}_e$ around its population counterpart $\Gamma_e := \mathrm{var}(\mathbb{E}[\boldsymbol{X} \mid Y])$ (see Lemma 2).

(iii) Based on the concentration inequality, we show that the space spanned by the top d eigenfunctions of $\widehat{\Gamma}_e$ is a root-n consistent estimator of the image of Γ_e (see Theorem 1). This part is a crucial step to our key results, the minimax rate optimality of FSIR estimator for the central space.

(iv) Having **(i)**-**(iii)** established, we apply the most widely used truncated scheme to estimate the inverse of the covariance operator of the predictor and then establish the consistency of the FSIR estimator for the central space. Furthermore, we identify the optimal truncation parameter to achieve the minimax optimal convergence rate for FISR in Theorem 2. It turns out that the converging rate of FSIR is the same as the minimax rate for estimating the slope in functional linear regression (Hall and Horowitz, 2007).

(v) Finally, we show that the convergence rate we obtained in **(iv)** is minimax rate-optimal for multiple-index models over a large class of distributions (see Theorem 3).

(vi) Simulation studies show that the optimal choice of m matches the theoretical ones and illustrate the efficiency of FSIR on both synthetic and real data.

To the best of our knowledge, this is the first work that rigorously establishes the optimality of FSIR for a general response Y. Our results provide a precise characterization of the difficulty associated with the estimation of the functional central space in terms of the minimax rates over a wide range of distributions. It not only enriches the existing theoretical results of FSIR, but also opens up new possibilities for extending other well-understood results derived from high-dimensional data to those related to functional data.

1.2 Notations and organization

Throughout the paper, we take $\mathcal{H} = L^2[0,1]$ to be the separable Hilbert space of square-integrable curves on $[0,1]$ with the inner product $\langle f, g\rangle = \int_0^1 f(u)g(u)\,\mathrm{d}u$ and norm $\|f\| := \sqrt{\langle f, f\rangle}$ for $f, g \in \mathcal{H}$.

For an operator T on $\mathcal{H}$, $\|T\|$ denotes its operator norm with respect to $\langle\cdot,\cdot\rangle$, i.e.,

$$\|T\| := \sup_{\boldsymbol{u}\in\mathbb{S}_{\mathcal{H}}} \|T(\boldsymbol{u})\|$$

where $\mathbb{S}_{\mathcal{H}} = \{\boldsymbol{u} \in \mathcal{H} : \|\boldsymbol{u}\| = 1\}$. $\mathrm{Im}(T)$ denotes the closure of the image of T, P_T the projection operator from $\mathcal{H}$ to $\mathrm{Im}(T)$, and T^* the adjoint operator of T (a bounded linear operator). If T is self-adjoint, $\lambda^+_{\min}(T)$ denotes the infimum of the positive spectrum of T and $T^\dagger$ the Moore–Penrose pseudo-inverse of T. Abusing notations, we also denote by P_S the projection operator onto a closed space $S \subseteq \mathcal{H}$. For any $x, y \in \mathcal{H}$, $x\otimes y$ is the operator of $\mathcal{H}$ to itself, defined by $x\otimes y(z) = \langle x, z\rangle y, \forall z \in \mathcal{H}$. For any random element $\boldsymbol{X} = \boldsymbol{X}_t \in \mathcal{H}$, its mean function is defined as $(\mathbb{E}\boldsymbol{X})_t = \mathbb{E}[\boldsymbol{X}_t]$. For any random operator T on $\mathcal{H}$, the mean $\mathbb{E}[T]$ is defined as the unique operator on $\mathcal{H}$ such that for all $z \in \mathcal{H}$, $(\mathbb{E}[T])(z) = \mathbb{E}[T(z)]$. Specifically, the covariance operator of $\boldsymbol{X}$, $\mathrm{var}(\boldsymbol{X})$, is defined as $\mathrm{var}(\boldsymbol{X})(z) = \mathbb{E}\left(\langle \boldsymbol{X}, z\rangle \boldsymbol{X}\right) - \langle \mathbb{E}\boldsymbol{X}, z\rangle \mathbb{E}\boldsymbol{X}$. For a pair of random variables $(\boldsymbol{X}, Y) \in \mathcal{H}\times\mathbb{R}$, Γ and Γ_e denote the covariance operator of

$\boldsymbol{X}$ and $\mathbb{E}[\boldsymbol{X} \mid Y]$ respectively, i.e.,

$$\Gamma := \text{var}(\boldsymbol{X}) \qquad \text{and} \qquad \Gamma_e := \text{var}(\mathbb{E}[\boldsymbol{X} \mid Y]). \tag{1.3}$$

For two sequences a_n and b_n, we denote $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$) if there exists a positive constant C such that $a_n \leqslant Cb_n$ (resp. $a_n \geqslant Cb_n$), respectively. We denote $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. For a random sequence X_n, we denote by $X_n = O_p(a_n)$ that $\forall \varepsilon > 0$, there exists a constant $C_\varepsilon > 0$, such that $\sup_n \mathbb{P}(|X_n| \geqslant C_\varepsilon a_n) \leqslant \varepsilon$. Let $[k]$ denote $\{1, 2, \ldots, k\}$ for some positive integer $k \geqslant 1$.

The rest of this paper is organized as follows. We first provide a brief review of FSIR in Section 2. After introducing the weak sliced stable condition for functional data in Section 3.1, we establish the root-n consistency of the estimated inverse regression subspace in Section 3.2. Lastly, the minimax rate optimality of FSIR are shown in Section 3.3 and the numerical experiments are reported in Section 4. All proofs are deferred to the supplementary files.

2. Optimal Truncated FSIR

Without loss of generality, we assume that $\boldsymbol{X} \in \mathcal{H}$ satisfies $\mathbb{E}[\boldsymbol{X}] = 0$ throughout the paper. As is usually done in functional data analysis (Ferré and Yao, 2003; Lian and Li, 2014; Lian, 2015), we assume that $\mathbb{E}[\|\boldsymbol{X}\|^4] < \infty$, which implies that Γ is a trace class (Hsing and Eubank, 2015) and $\boldsymbol{X}$ possesses the following Karhunen–Loéve expansion:

$$\boldsymbol{X} = \sum_{i=1}^{\infty} \xi_i \phi_i \tag{2.1}$$

where ξ_i's are random variables satisfying $\mathbb{E}[\xi_i^2] = \lambda_i$ and $\mathbb{E}[\xi_i \xi_j] = 0$ for $i \neq j$ and $\{\phi_i\}_{i=1}^{\infty}$ are the eigenfunctions of Γ in (1.3) associated with the decreasing eigenvalues sequence $\{\lambda_i\}_{i=1}^{\infty}$. In addition, we assume that Γ is non-singular (i.e., $\lambda_i > 0, \forall i$) as the literature on functional data analysis usually does. Since Γ is compact (Γ is a trace class), by spectral decomposition theorem of compact operators, we know that $\{\phi_i\}_{i=1}^{\infty}$ forms a complete basis of $\mathcal{H}$.

In order for FSIR to produce a consistent estimator of the functional central space $\mathcal{S}_{Y|\boldsymbol{X}}$ for $(\boldsymbol{X}, Y)$ from the multiple index model (1.2), people often assume that the joint distribution of $(\boldsymbol{X}, Y)$ satisfies the following conditions (see e.g., Ferré and Yao (2003); Lian and Li (2014); Lian (2015)).

Assumption 1. The joint distribution of $(\boldsymbol{X}, Y)$ satisfies

i) *Linearity condition:* For any $\boldsymbol{b} \in \mathcal{H}$, $\mathbb{E}[\langle \boldsymbol{b}, \boldsymbol{X} \rangle \mid (\langle \boldsymbol{\beta}_1, \boldsymbol{X} \rangle, \ldots, \langle \boldsymbol{\beta}_d, \boldsymbol{X} \rangle)]$ is linear in $\langle \boldsymbol{\beta}_1, \boldsymbol{X} \rangle, \ldots, \langle \boldsymbol{\beta}_d, \boldsymbol{X} \rangle$.

ii) *Coverage condition:* $\mathrm{Rank}\left(\mathrm{var}(\mathbb{E}[\boldsymbol{X}|Y])\right) = d$.

Both of these conditions are natural generalizations of the multivariate ones that appear in the multivariate SIR literature (Li, 1991; Hall and Li, 1993; Li and Hsing, 2010). They are necessary for Ferré and Yao (2003) to establish that the *inverse regression subspace* $\mathcal{S}_e := \mathrm{span}\{\mathbb{E}[\boldsymbol{X} \mid Y = y] \mid y \in \mathbb{R}\}$ equals to the space $\Gamma \mathcal{S}_{Y|\boldsymbol{X}} := \mathrm{span}\{\Gamma \boldsymbol{\beta}_1, \ldots, \Gamma \boldsymbol{\beta}_d\}$. Since $\mathcal{S}_e = \mathrm{Im}(\Gamma_e)$, FSIR estimates $\mathcal{S}_{Y|\boldsymbol{X}}$ by estimating $\Gamma^{-1}\mathrm{Im}(\Gamma_e)$.

The FSIR procedure for estimating Γ_e can be briefly summarized as follows. Given n i.i.d. samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$ from the multiple index model (1.2), FSIR sorts the samples according to the order statistics $Y_{(i)}$ and then divide the samples into $H(\geqslant d)$ equal-size slices (for the simplicity of notation, we assume that $n = Hc$ for some positive integer c).

We re-index the data as

$$Y_{h,j} = Y_{(c(h-1)+j)} \qquad \text{and} \qquad \boldsymbol{X}_{h,j} = \boldsymbol{X}_{(c(h-1)+j)}$$

where $\boldsymbol{X}_{(k)}$ is the concomitant of $Y_{(k)}$ (Yang, 1977). Let $\mathcal{S}_h$ be the h-th interval $(Y_{(h-1,c)}, Y_{(h,c)}]$ for $h = 2, \ldots, H-1$, $\mathcal{S}_1 = \{y \mid y \leqslant Y_{(1,c)}\}$ and $\mathcal{S}_H = \{y \mid y > Y_{(H-1,c)}\}$. Consequently, $\mathfrak{S}_H(n) := \{\mathcal{S}_h, h = 1, .., H\}$ is a partition of $\mathbb{R}$ and is referred to as the *sliced partition*. FSIR estimates the conditional covariance Γ_e via

$$\widehat{\Gamma}_e := \frac{1}{H} \sum_{h=1}^{H} \overline{\boldsymbol{X}}_{h,\cdot} \otimes \overline{\boldsymbol{X}}_{h,\cdot} \tag{2.2}$$

where $\overline{\boldsymbol{X}}_{h,\cdot} := \frac{1}{c} \sum_{j=1}^{c} \boldsymbol{X}_{h,j}$ is the sample mean in the h-th slice.

To estimate Γ^{-1}, one need to resort to some truncation scheme. Given n i.i.d. samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$, a straightforward estimator of Γ^{-1} is $\widehat{\Gamma}^{\dagger}$, the pseudo-inverse of the sample covariance operator $\widehat{\Gamma} := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \otimes \boldsymbol{X}_i$. However, it is not practical since the operator Γ is compact (Γ is a trace class) and then Γ^{-1} is unbounded. To circumvent this technical difficulty, one may apply some truncation strategies such as the operations in Ferré and Yao (2003), which we briefly review as follows. We choose an integer m and define the truncated covariance operator $\Gamma_m := \Pi_m \Gamma \Pi_m$ where $\Pi_m := \sum_{i=1}^{m} \phi_i \otimes \phi_i$ is the truncation projection operator. Since each Γ_m is of finite rank, we are able to estimate $\Gamma_m^{\dagger}$. Specifically, let the sample truncation operator $\widehat{\Pi}_m := \sum_{i=1}^{m} \widehat{\phi}_i \otimes \widehat{\phi}_i$ and the sample truncated covariance operator $\widehat{\Gamma}_m := \widehat{\Pi}_m \widehat{\Gamma} \widehat{\Pi}_m$, where $\{\widehat{\phi}_m\}_{i=1}^{m}$ are the top m eigenfunctions of $\widehat{\Gamma}$. Then the estimator of $\Gamma_m^{\dagger}$ can be defined as $\widehat{\Gamma}_m^{\dagger}$. It is clear that $\|\Gamma - \Gamma_m\| \xrightarrow{m \to \infty} 0$ and the space $\Gamma_m \mathcal{S}_{Y|\boldsymbol{X}}$ would be close to the space $\Gamma \mathcal{S}_{Y|\boldsymbol{X}}$ when m is sufficiently large. Thus we can accurately estimate Γ^{-1}

by $\widehat{\Gamma}_m^{\dagger}$ for sufficiently large m.

Algorithm 1 FSIR (Ferré and Yao, 2003).

1. Standardize $\{\boldsymbol{X}_i\}_{i=1}^n$, i.e., $\boldsymbol{Z}_i := \boldsymbol{X}_i - n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$;
2. Divide the n samples $\{(\boldsymbol{Z}_i, Y_i)\}_{i=1}^n$ into H equally sized slices according to the order statistics $Y_{(i)}, 1 \leqslant i \leqslant n$;
3. Calculate $\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i \otimes \boldsymbol{Z}_i$ with its top m eigenvalues $\{\widehat{\lambda}_i, 1 \leqslant i \leqslant m\}$ and the corresponding eigenfunctions $\{\widehat{\phi}_i : 1 \leqslant i \leqslant m\}$, where m is the tuning parameter. Let $\widehat{\Gamma}_m = \sum_{i=1}^m \widehat{\lambda}_i \widehat{\phi}_i \otimes \widehat{\phi}_i$;
4. Calculate $\overline{\boldsymbol{Z}}_{h,\cdot} = \frac{1}{c}\sum_{j=1}^c \boldsymbol{Z}_{h,j}$, $h = 1, 2, \cdots, H$ and $\widehat{\Gamma}_e = \frac{1}{H}\sum_{h=1}^H \overline{\boldsymbol{Z}}_{h,\cdot} \otimes \overline{\boldsymbol{Z}}_{h,\cdot}$ similarly as (2.2) ;
5. Find the top d eigenfunctions of $\widehat{\Gamma}_e$, denoted by $\widehat{\boldsymbol{\beta}}_k'$ $(k = 1, \ldots, d)$ and calculate $\widehat{\boldsymbol{\beta}}_k = \widehat{\Gamma}_m^{\dagger}\widehat{\boldsymbol{\beta}}_k'$.

Return span $\left\{\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_d\right\}$.

In introducing our *optimal truncated FSIR algorithm* (FSIR-OT), we commence by revisiting the classical FSIR algorithm proposed in Ferré and Yao (2003) as shown in Algorithm1. It is worth noting that Ferré and Yao (2003) did not provide any specific guidance on the choice of the tuning parameter m. Our FSIR-OT algorithm provides an optimal selection criterion for m, namely, $m \propto n^{1/(\alpha+2\beta)}$, where α and β are defined in Assumption 4. Under this optimal choice, we prove that FSIR-OT can achieve the minimax rate for estimating the central space in the next section.

3. Minimax rate optimality of FSIR-OT

Throughout the paper, the number of indexes d is assumed known and fixed. By analyzing the asymptotic behaviors of FSIR-OT, we derive the minimax rate optimality of FSIR-

OT. We begin by proposing a fairly mild condition called the weak sliced stable condition (WSSC) for functional data. Then, we show that the top d eigenfunctions of the estimated conditional covariance $\widehat{\Gamma}_e$ span a consistent estimator $\widehat{\mathcal{S}}_e$ of the inverse regression subspace $\mathcal{S}_e$, with convergence rate of $n^{-1/2}$ based on WSSC. Lastly, we establish the consistency of the FSIR-OT estimator of the central space and show that the convergence rate is minimax optimal over a large class of distributions.

3.1 Weak sliced stable condition for functional data

The sliced stable condition (SSC) was first introduced in Lin et al. (2018) to analyze the asymptotic behavior of SIR in high dimensions such as the phase transition phenomenon. Lin et al. (2021) showed the optimality of SIR in high dimensions based on SSC. Huang et al. (2023) weakened SSC to weak sliced stable condition (WSSC) to establish the optimality of SIR in more general settings. This inspires us to extend WSSC to functional data. Throughout the paper, γ is a fixed small positive constant.

Definition 1 (**Weak Sliced Stable Condition**)**.** Let $Y \in \mathbb{R}$ be a random variable, K a positive integer and $\tau > 1$ a constant. A partition $\mathcal{B}_H := \{-\infty = a_0 < a_1 < \cdots < a_{H-1} < a_H = \infty\}$ of $\mathbb{R}$ is called a γ-partition if

$$\frac{1-\gamma}{H} \leqslant \mathbb{P}(a_h \leqslant Y \leqslant a_{h+1}) \leqslant \frac{1+\gamma}{H}, \qquad \forall h = 0, 1, \ldots, H-1. \tag{3.1}$$

A continuous curve $\boldsymbol{\kappa}(y) : \mathbb{R} \to \mathcal{H}$ is said to be *weak* (K, τ)*-sliced stable* w.r.t. Y, if for any

$H \geqslant K$ and any γ-partition $\mathcal{B}_H$, it holds that

$$\frac{1}{H}\sum_{h=0}^{H-1} \operatorname{var}\left(\langle \boldsymbol{u}, \boldsymbol{\kappa}(Y)\rangle \mid a_h \leqslant Y \leqslant a_{h+1}\right) \leqslant \frac{1}{\tau}\operatorname{var}\left(\langle \boldsymbol{u}, \boldsymbol{\kappa}(Y)\rangle\right) \quad (\forall \boldsymbol{u} \in \mathbb{S}_{\mathcal{H}}). \tag{3.2}$$

Compared with the original SSC (e.g., the equation (4) in Lin et al. (2018)) for the central curve $\boldsymbol{m}(y) := \mathbb{E}[\boldsymbol{X}|Y = y]$, WSSC condition is less restrictive. The average of the variances (the left hand side of (3.2)) is only required to be sufficiently small by WSSC condition, in contrast to that, it needs to vanish as $H \to \infty$ by the original SSC. In fact, as we will in Theorem 1, the constant τ in (3.2) only needs to be greater than $\frac{6\|\Gamma_e\|}{\lambda^{+}_{\min}(\Gamma_e)}$ to guarantee the consistency of the FSIR-OT estimator. Furthermore, the following lemma shows that WSSC of $\boldsymbol{m}(y)$ is readily fulfilled under certain mild prerequisites.

Lemma 1. *Suppose that the joint distribution of* $(\boldsymbol{X}, Y) \in \mathcal{H} \times \mathbb{R}$ *satisfies the following conditions:*

i) *for any* $\boldsymbol{u} \in \mathbb{S}_{\mathcal{H}}$, $\mathbb{E}\left[|\langle \boldsymbol{u}, \boldsymbol{X}\rangle|^{\ell}\right] \leqslant c_1$ *holds for absolute constants* $\ell > 2$ *and* $c_1 > 0$*;*

ii) Y *is a continuous random variable;*

iii) *the central curve* $\boldsymbol{m}(y) := \mathbb{E}[\boldsymbol{X}|Y = y]$ *is continuous.*

Then for any $\tau > 1$*, there exists an integer* $K = K(\tau, d) \geqslant d$ *such that* $\boldsymbol{m}(y)$ *is weak* (K, τ)*-sliced stable w.r.t.* Y*.*

Assumption 2. The central curve $\boldsymbol{m}(y) = \mathbb{E}[\boldsymbol{X}|Y = y]$ is weak (K, τ)-sliced stable with respect to Y for two positive constants K and τ (i.e., WSSC).

We note that the requirement of K being a constant is mild since d is bounded. With the help of WSSC, we can now bound the distance between $\widetilde{\Gamma}_e$ and Γ_e, where

$$\widetilde{\Gamma}_e := \frac{1}{H} \sum_{h:\mathcal{S}_h \in \mathfrak{S}_H(n)} \overline{\boldsymbol{m}}_h \otimes \overline{\boldsymbol{m}}_h \quad \text{and} \quad \overline{\boldsymbol{m}}_h := \mathbb{E}[\boldsymbol{m}(Y) \mid Y \in \mathcal{S}_h] = \mathbb{E}[\boldsymbol{X} | Y \in \mathcal{S}_h].$$

Here $\mathfrak{S}_H(n)$ is the sliced partition defined in Section 2. This bound is key to obtaining a concentration inequality for the FSIR-OT estimator $\widehat{\Gamma}_e$ of the conditional covariance Γ_e.

Proposition 1. *Under Assumption 2, there exist positive constants C and $H_0 \geqslant K$, such that for all $H > H_0$, if $n > 1 + 4H/\gamma$ is sufficiently large, we have*

$$\mathbb{P}\left(\left|\left\langle \left(\widetilde{\Gamma}_e - \Gamma_e\right)(\boldsymbol{u}), \boldsymbol{u}\right\rangle\right| \leqslant \frac{3}{\tau}\left\langle \Gamma_e(\boldsymbol{u}), \boldsymbol{u}\right\rangle, \forall \boldsymbol{u} \in \mathbb{S}_{\mathcal{H}}\right) \geqslant 1 - CH^2\sqrt{n+1}\exp\left(\frac{-\gamma^2(n+1)}{32H^2}\right). \tag{3.3}$$

When $\tau > \frac{6\|\Gamma_e\|}{\lambda^+_{\min}(\Gamma_e)}$, we know that $\mathrm{Im}(\widetilde{\Gamma}_e) = \mathrm{Im}(\Gamma_e)$ holds with high probability (see Lemma 5 in Appendix for details). In other words, Proposition 1 implies that $\mathrm{Im}(\widetilde{\Gamma}_e)$ is a consistent estimator of $\mathrm{Im}(\Gamma_e)$ even if $\widetilde{\Gamma}_e$ is not a consistent estimator of Γ_e.

3.2 Root-n consistency of the FSIR-OT estimator for inverse regression subspace

We study asymptotic behaviors of the FSIR-OT estimator $\widehat{\mathcal{S}}_e$ of the inverse regression subspace $\mathcal{S}_e$. As in most studies in functional data analysis (Hall and Horowitz, 2007; Lei, 2014; Lian, 2015; Wang and Lian, 2020), we introduce the following assumption:

Assumption 3. There exists a constant $c_1 > 0$ such that $\mathbb{E}[\xi_i^4]/\lambda_i^2 \leqslant c_1$ uniformly for all $i \in \mathbb{Z}_+$ where ξ_i and λ_i are defined in (2.1).

Now we are ready to state our first main result, which is similar to the 'key lemma' in Lin et al. (2018), a crucial tool for developing the phase transition phenomenon and establishing the minimax optimality of the high dimensional SIR.

Lemma 2. *Suppose that Assumptions 2 and 3 hold. For any fixed integer $H > H_0$ (H_0 is defined in Proposition 1) and any sufficiently large $n > 1 + 4H/\gamma$, we have*

$$\left\|\widehat{\Gamma}_e - \Gamma_e\right\| = O_p\left(\frac{1}{\tau} + \sqrt{\frac{1}{n}}\right) \quad \text{and} \quad \left\|\widehat{\Gamma}_e - \widetilde{\Gamma}_e\right\| = O_p\left(\sqrt{\frac{1}{n}}\right).$$

The τ term in the first equation of Lemma 2 suggests that $\widehat{\Gamma}_e$ may not be a consistent estimator of Γ_e. However, we are interested in estimating the space $\mathcal{S}_e = \mathrm{Im}(\Gamma_e)$ rather than Γ_e itself and the τ term would not affect the convergence rate of $\left\|P_{\widehat{\mathcal{S}}_e} - P_{\mathcal{S}_e}\right\|$ as long as τ is sufficiently large. This will be elaborated in the following theorem, our second main result.

Theorem 1. *Consider the same conditions and constants as in Lemma 2 and suppose that $\tau > \frac{6\|\Gamma_e\|}{\lambda^+_{\min}(\Gamma_e)}$. It holds that*

$$\mathbb{E}\left[\left\|P_{\widehat{\mathcal{S}}_e} - P_{\mathcal{S}_e}\right\|^2\right] \lesssim \frac{1}{n} \tag{3.4}$$

where the expectation $\mathbb{E}$ is taken with respect to the randomness of the sample.

Equation (3.4) implies that $\widehat{\mathcal{S}}_e$ is a root-n consistent estimator of the inverse regression subspace $\mathcal{S}_e$. This is a crucial step to establish the minimax rate optimality of FSIR-OT estimator for the central space.

3.3 Optimality of FSIR-OT

In order to obtain the convergence rate of the FSIR-OT estimator of the central space, we need a further assumption, which is commonly imposed in functional data analysis (see e.g., Hall and Horowitz (2007); Lei (2014); Lian (2015)).

Assumption 4 (Rate-type condition)**.** There exist positive constants α, β, c_2 and c_2' satisfying

$$\alpha > 1, \quad \frac{1}{2}\alpha + 1 < \beta, \quad \lambda_j - \lambda_{j+1} \geqslant c_2 j^{-\alpha-1} \text{ and } |b_{ij}| \leqslant c_2' j^{-\beta} \quad (\forall i \in [d], j \in \mathbb{Z}_+)$$

where $b_{ij} := \langle \eta_i, \phi_j \rangle$ for $\{\eta_i\}_{i=1}^d$ the generalized eigenfunctions of Γ_e associated with top d eigenvalues $\{\mu_i\}_{i=1}^d$ (i.e., $\Gamma_e \eta_i = \mu_i \Gamma \eta_i$).

The assumption on the eigenvalues λ_j of Γ requires a gap between adjacent eigenvalues and ensures the accuracy of the estimation of eigenfunctions of Γ. It also implies a lower bound on the decay rate of λ_j: $\lambda_j \gtrsim j^{-\alpha}$. The assumption on the coefficients b_{ij} implies that they do not decrease too slowly with respect to j uniformly for all i. It also implies that any basis $\{\widetilde{\boldsymbol{\beta}}_i\}_{i=1}^d$ of $\mathcal{S}_{Y|\boldsymbol{X}}$ such that $\widetilde{\boldsymbol{\beta}}_i = \sum_{j=1}^\infty \widetilde{b}_{ij}\phi_j$ satisfies $|\widetilde{b}_{ij}| \lesssim j^{-\beta}$. The inequality $\frac{1}{2}\alpha + 1 < \beta$ requires that the generalized eigenfunction η_i is smoother than the covariate function $\boldsymbol{X}$.

The conditions in Assumption 4 have been imposed in Hall and Horowitz (2007) for showing that the minimax rate of functional linear regression models is $n^{-(2\beta-1)/(\alpha+2\beta)}$. Lian (2015) also made use of some similar conditions to show that the FSIR estimator of the central space $\mathcal{S}_{Y|\boldsymbol{X}}$ for discrete Y (i.e., Y only takes finite values) can achieve the same convergence rate as the one for estimating the slope in functional linear regression.

Now we state our third main result, an upper bound on the convergence rate of the FSIR-OT estimator of the central space.

Theorem 2. *Suppose Assumptions 1 to 4 hold with constants α, β and $\tau > \frac{6\|\Gamma_e\|}{\lambda^+_{\min}(\Gamma_e)}$. By choosing $m \asymp n^{\frac{1}{\alpha+2\beta}}$, we can get that for any fixed integer $H > H_0$ (H_0 is defined in*

Proposition 1) and any sufficiently large $n > 1 + 4H/\gamma$*, we have*

$$\left\| P_{\widehat{\mathcal{S}}_{Y|\boldsymbol{X}}} - P_{\mathcal{S}_{Y|\boldsymbol{X}}} \right\|^2 = O_p\left(n^{\frac{-(2\beta-1)}{\alpha+2\beta}}\right)$$

where $\widehat{\mathcal{S}}_{Y|\boldsymbol{X}} = \widehat{\Gamma}_m^{\dagger} \widehat{\mathcal{S}}_e$ *is the estimated central space given by FSIR-OT.*

The convergence rate we have derived for FSIR-OT is the same as the minimax rate for estimating the slope in functional linear regression (Hall and Horowitz, 2007). While the convergence rate appears to be the same as that in Lian (2015), their study only considered the case where the response Y is discrete. Moreover, their work lacked a proof for the optimality of FSIR-OT in estimating the central space. Yao et al. (2015) also introduced the FCSE method, focusing on scenarios with sparse designs, wherein only limited, noisy, and irregular observations are available for some or all subjects. However, they did not provide any analysis regarding the minimax optimality. In the following, we will provide a rigorous proof that our convergence rate is indeed minimax rate-optimal over a large class of distributions, which is highly nontrivial. To do this, we first introduce a class of distributions:

$$\mathfrak{M}(\alpha, \beta, \tau, c_0, C_0) := \left\{ (\boldsymbol{X}, Y) \;\middle|\; \begin{array}{l} Y = f(\langle \boldsymbol{\beta}_1, \boldsymbol{X} \rangle, \ldots, \langle \boldsymbol{\beta}_d, \boldsymbol{X} \rangle, \varepsilon); \\ \boldsymbol{X}, \boldsymbol{\beta}_i \in \mathcal{H} := L^2[0,1] \quad (i = 1, \ldots, d); \\ \varepsilon \text{ is a random noise independent of } \boldsymbol{X}; \\ (\boldsymbol{X}, Y) \text{ satisfies Assumption 1-4 }; \\ c_0 \leqslant \lambda_d(\Gamma_e) \leqslant \cdots \leqslant \lambda_1(\Gamma_e) \leqslant C_0; \\ \|\Gamma\| \leqslant C_0, \quad \lambda_{\min}(\Gamma|_{\mathcal{S}_e}) \geqslant c_0 \end{array} \right\}$$

where c_0 and C_0 are two positive universal constants.

Then we have the following minimax lower bound for estimating the central space over $\mathfrak{M}(\alpha,\beta,\tau,c_0,C_0)$.

Theorem 3. *For any given positive constants α, β and τ satisfying $\alpha > 1, \frac{1}{2}\alpha + 1 < \beta$ and $\tau > \frac{6C_0}{c_0}$, there exists an absolute constant $\vartheta > 0$ that only depends on α and β, such that for any sufficiently large n, it holds that*

$$\inf_{\widehat{\mathcal{S}}_{Y|\boldsymbol{X}}} \sup_{\mathcal{M}\in\mathfrak{M}(\alpha,\beta,\tau,c_0,C_0)} \mathbb{P}_{\mathcal{M}}\left(\left\|P_{\widehat{\mathcal{S}}_{Y|\boldsymbol{X}}} - P_{\mathcal{S}_{Y|\boldsymbol{X}}}\right\|^2 \geqslant \vartheta n^{-\frac{2\beta-1}{\alpha+2\beta}}\right) \geqslant 0.9$$

where $\widehat{\mathcal{S}}_{Y|\boldsymbol{X}}$ is taken over all possible estimators of $\mathcal{S}_{Y|\boldsymbol{X}}$ based on the training data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$.

The main tool we used in proving this minimax lower bound is Fano's Lemma (see e.g., (Yu, 1997)). The major challenge is to construct a specific family of distributions that are far apart from each other in the parameter space, yet close to each other in terms of Kullback–Leibler divergence. An important contribution in this paper is the construction of such distributions.

Theorems 2 and 3 together show that the FSIR-OT estimator $P_{\widehat{\mathcal{S}}_{Y|\boldsymbol{X}}}$ is minimax rate-optimal for estimating the central space.

4. Numerical Studies

In this section, we present several numerical experiments to illustrate the behavior of the FSIR-OT algorithm. The first experiment demonstrates the optimal choice of the truncation parameter m for estimating the central space. The results corroborate the conclusion of Theorem 2 that the choice in of FSIR-OT (namely $m \asymp n^{\frac{1}{\alpha+2\beta}}$) is optimal. The second experiment focuses on the estimation performance of FSIR-OT on synthetic data. Lastly, we analyze a real data set on bike rentals using FSIR algorithms. By comparing our algorithm with the FCSE algorithm of Yao et al. (2015) and the regularized FSIR (Lian 2015, RFSIR), we demonstrate advantages of FSIR-OT on both synthetic and real datasets. Similar to FSIR-OT, FCSE performs a truncation operation on the covariance operator, controlled by parameter m, whereas RFSIR employs ridge-type regularization characterized by a regularization parameter ρ.

4.1 Generalized signal noise ratio (gSNR) of multiple index models

Recall that the signal-to-noise ratio (SNR) for the linear model $Y = \langle \boldsymbol{\beta}, \boldsymbol{X} \rangle + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, is defined as

$$\mathrm{SNR} = \frac{\mathbb{E}[\langle \boldsymbol{\beta}, \boldsymbol{X} \rangle^2]}{\mathbb{E}[Y^2]} = \frac{\langle \Gamma\boldsymbol{\beta}, \boldsymbol{\beta} \rangle}{\sigma^2 + \langle \Gamma\boldsymbol{\beta}, \boldsymbol{\beta} \rangle}.$$

A simple calculation shows that

$$\Gamma_e = \frac{\Gamma\boldsymbol{\beta} \otimes \Gamma\boldsymbol{\beta}}{\langle \Gamma\boldsymbol{\beta}, \boldsymbol{\beta} \rangle + \sigma^2}, \quad \text{and} \quad \lambda(\Gamma_e) = \frac{\|\Gamma\boldsymbol{\beta}\|^2}{\langle \Gamma\boldsymbol{\beta}, \boldsymbol{\beta} \rangle + \sigma^2},$$

where $\lambda(\Gamma_e)$ is the unique non-zero eigenvalue of Γ_e. This leads to the following identity for the linear model:

$$\lambda(\Gamma_e) = \frac{\|\Gamma\boldsymbol{\beta}\|^2}{\langle \Gamma\boldsymbol{\beta}, \boldsymbol{\beta} \rangle}\mathrm{SNR}.$$

Thus, in a multiple index model we call λ, the smallest non-zero eigenvalue of Γ_e, the model's generalized SNR (gSNR) .

4.2 Optimal choice of truncation parameter m

Throughout this section, we set $H = 15$ and $\varepsilon \sim N(0, 2)$. We note that the results are not sensitive to the choice of H. The guidelines for the choice of H in practice are presented Section H.2 of Supplementary Material. The experimental results for other noise levels (with variances of 1 and 0.25, respectively) and other H are shown in Section H.3 of Supplementary Material.

The following model is first considered:

(I) $Y = \langle \boldsymbol{\beta}_1, \boldsymbol{X} \rangle + \varepsilon$, where $\boldsymbol{X} = \sum_{j=1}^{100} j^{-3/4} X_j \phi_j$ and $\boldsymbol{\beta}_1 = \sum_{j \geqslant 1} (-1)^j j^{-2} \phi_j$. Here $X_j \overset{\text{iid}}{\sim} N(0, 1)$, $\phi_1 = 1, \phi_{j+1} = \sqrt{2}\cos(j\pi t), j \geqslant 1$.

Note that the construction of $\boldsymbol{X}$ here is equivalent to a construction that satisfies the assumption that Γ is non-singular (i.e., $\lambda_i > 0, \forall i$). A detailed explanation is deferred to Section H.1 of Supplementary Material.

For this model $\alpha = 3/2$ and $\beta = 2$, so the optimal choice of m used by FSIR-OT satisfies $m \propto n^{2/11}$. The gSNRs of Model I are 0.791, 0.498, and 0.333, respectively, when the noise variances are 0.25, 1, and 2.

To evaluate the performance of FSIR-OT, we consider the subspace estimation error defined as $\mathcal{D}(\widehat{\boldsymbol{B}}; \boldsymbol{B}) := \left\| P_{\widehat{\boldsymbol{B}}} - P_{\boldsymbol{B}} \right\|$ where $\widehat{\boldsymbol{B}} := (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_d) : \mathbb{R}^d \to L^2[0, 1]$ and $\boldsymbol{B} := (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) : \mathbb{R}^d \to L^2[0, 1]$. This metric takes value in $[0, 1]$ and, the smaller it is, the better the performance. Each trial is repeated 100 times for reliability.

The left panel of Figure 1 is the average subspace estimation error under Model (I) where n ranges in $\{2 \times 10^3, 2 \times 10^4, 5 \times 10^4, 2 \times 10^5, 5 \times 10^5, 10^6\}$, m ranges in $\{3, 4, \ldots, 25\}$. The optimal value of m (denoted by m^*) for each n is marked with a red circle. Among the 100 replicates for every n, the number of times that the minimal estimation error occurs at m^* is $48, 32, 29, 29, 26, 26$, respectively. The shaded areas represent the standard error bands associated with these estimates (all smaller than 0.009). The right panel of Figure 1 illustrates the linear dependence of $\log(m^*)$ on $\log(n)$. The solid line characterizes the linear trend of $\log(m^*)$ against $\log(n)$. The dotted line is their least-squares fitting, with its slope estimated as 0.2, which is close to the theoretical value of $2/11$. These results are consistent with the theoretically optimal choice of m in FSIR-OT.

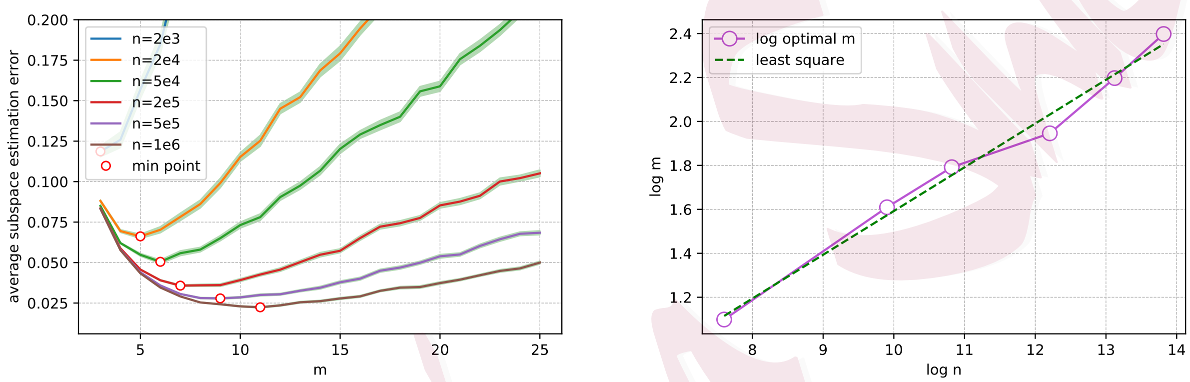

Figure 1: Experiments for the optimal choice of truncation parameter m with $\varepsilon \sim N(0, 2)$ and $H = 15$. Left: average subspace estimation error with increasing m for different n. Right: linear trend of $\log(m^*)$ against $\log(n)$, with a slope of 0.2 and $R^2 > 0.98$.

4.3 Subspace estimation error performance in synthetic data

In this section, we compare FSIR-OT with RFSIR and FCSE for model (I) from Section 4.2 and the following two models:

(II) $Y = \langle \boldsymbol{\beta}_1, \boldsymbol{X} \rangle + 100\langle \boldsymbol{\beta}_2, \boldsymbol{X} \rangle^3 + \varepsilon$, where $\boldsymbol{\beta}_1(t) = \sqrt{2}\sin(\frac{3\pi t}{2})$, $\boldsymbol{\beta}_2(t) = \sqrt{2}\sin(\frac{5\pi t}{2})$ for $t \in [0, 1]$, and $\boldsymbol{X}$ is the standard Brownian motion on $[0, 1]$ (The Brownian motion is approximated by the top 100 eigenfunctions of the Karhunen–Loève decomposition in practical implementation).

(III) $Y = \exp(\langle \boldsymbol{\beta}, \boldsymbol{X} \rangle) + \varepsilon$, where $\boldsymbol{X}$ is the standard Brownian motion on $[0, 1]$, and $\boldsymbol{\beta} =$

$\sqrt{2}\sin(\frac{3\pi t}{2})$.

For model II and model III, we compute the estimated gSNR by $\lambda_d(\widehat{\Gamma}_e)$, the d-th eigenvalue of the SIR estimate of Γ_e based on 2000 replicates, where $n = 10000$. The mean gSNRs (standard deviation) of model II are 0.020 (0.001), 0.009 (0.001), and 0.003 (0.001), respectively, when the noise variances are 0.25, 1, and 2. The mean gSNRs (standard deviation) of model III are 0.729 (0.01), 0.536 (0.01), and 0.305 (0.01), respectively, when the noise variances are 0.25, 1, and 2.

For each model, we calculate the average subspace estimation error of FSIR-OT, RFSIR and FCSE based on 100 replicates, where $n = 20000$, the truncation parameter m of FSIR-OT and FCSE ranges in $\{2, 3, \ldots, 13, 14, 20, 30, 40\}$, and the regularization parameter ρ in RFSIR ranges in $0.01 \times \{1, 2, \cdots, 9, 10, 15, 20, 25, 30, 40, \cdots, 140, 150\}$. Detailed results are presented in Figure 2, where we mark the minimal error in each model with red '×' and denote the corresponding value of truncation (or regularization) parameter by m^* (or ρ^*). The shaded areas represent the corresponding standard errors, all of which are less than 0.012. For FSIR-OT, the minimal errors for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are $0.06, 0.03$, and 0.01 respectively. Among the 100 replicates for every model, the number of times that the minimal estimation error occurs at m^* is $33, 80$ and 70, respectively. For RFSIR, the corresponding minimal errors are 0.10, 0.08, and 0.01, respectively. Among the 100 replicates for every model, the number of times that the minimal estimation error occurs at ρ^* is $37, 13$ and 18, respectively. For FCSE, the corresponding minimal errors are $0.07, 0.03$, and 0.02. Among the 100 replicates for every model, the number of times that the minimal estimation error occurs at m^* is $24, 23$ and 28, respectively.

The results here suggest that the performance of FSIR-OT is generally superior to, or at the very least equivalent to, that of RFSIR and FCSE.

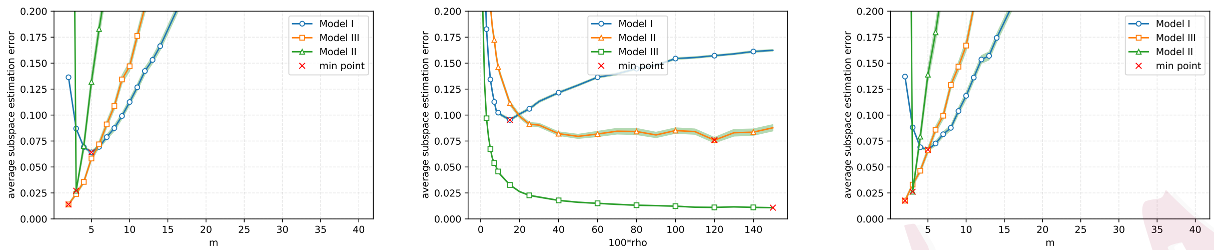

Figure 2: Average subspace estimation error of FSIR-OT, RFSIR and FCSE for various models in the case of $\varepsilon \sim N(0, 2)$ and $H = 15$. The standard errors are all below 0.01. Left: FSIR-OT with different truncation parameter m; Middle: RFSIR with different values of the regularization parameter ρ; Right: FCSE with different truncation parameter m.

4.4 Application to real data

In the following, we apply FSIR-OT to a business data analysis problem regarding bike sharing. The data are available from `https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`. The main purpose is to analyze how the bike rental counts are affected by the temperature on Saturdays. After removing data from 3 Saturdays with missing information, we plot hourly bike rental counts and hourly normalized temperature (values divided by the maximum 41°C) on 102 Saturdays in Figure 3. In the following experiments, we treat hourly normalized temperature and the logarithm of daily average bike rental counts as predictor function and scalar response respectively.

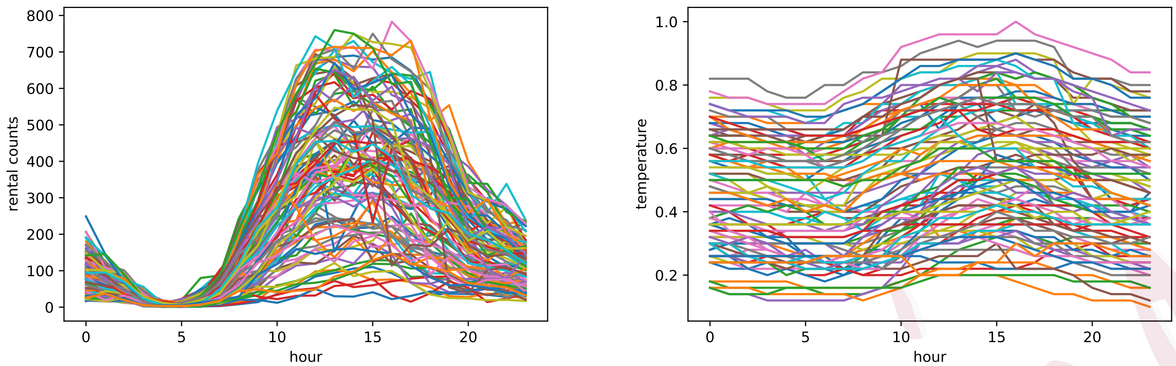

Figure 3: Bike sharing data

In order to compare the estimation error performance of FSIR-OT with RFSIR and FCSE for estimating the central space, we employ dimension reduction using these algorithms with $H = 15$ as an intermediate step in modelling the relation between the predictor and response. Specifically, given any training samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, we utilize each dimension reduction algorithm to obtain a set of low-dimensional predictors $\boldsymbol{x}_i$ for $i \in [n]$. Then, we employ Gaussian process regression to fit a nonparametric regression model based on samples $\{(\boldsymbol{x}_i, Y_i)\}_{i=1}^n$. We randomly select 90 samples as the training data and then calculate the out-of-sample mean squared error (MSE) using the remaining samples.

Since d is unknown in most real applications, including this one, we follow the PCA approach by calculating the sum of the first k ($k \leq 24$) eigenvalues of $\widehat{\Gamma}_e$. We found for this dataset that the first five eigenvalues account for 99.8% of the summation of all eigenvalues. Therefore, we narrowed the selection range of d to $\{1, 2, 3, 4, 5\}$. For each chosen d, we then selected the corresponding value of m satisfying $m \geq d$. The experiment is repeated 100 times and the mean and standard error are presented in Table 1. Among the 100 replicates for every method (FSIR-OT, FCSE, and RFSIR), the number of times that the minimal estimation error occurs at the optimal pair of tuning parameters (i.e., (m^*, d^*), (m^*, d^*), and

(ρ^*, d^*), respectively) is 6, 4, and 8, respectively.

From Table 1, it can be concluded that FSIR-OT performs better than FCSE and RFSIR if both methods are fine-tuned. Furthermore, the best result of FSIR-OT is observed at $d = 2$, while those for FCSE and RFSIR are both observed at $d = 4$. This means that, if all methods were further fine-tuned, FSIR-OT would have provided a more accurate and simpler (lower dimensional) model for the relationship between the response variable and the predictor than the other two methods.

$H = 15$	m	2	4	6	8	10
FSIR-OT	$d = 1$	**0.209** (0.010)	**0.212** (0.010)	**0.207** (0.010)	**0.211** (0.009)	**0.206** (0.009)
	$d = 2$	**0.229** (0.013)	**0.196** (0.009)	**0.188** (0.008)	**0.200** (0.010)	**0.215** (0.010)
	$d = 3$		**0.209** (0.012)	**0.193** (0.009)	**0.208** (0.010)	**0.207** (0.010)
	$d = 4$		**0.229** (0.011)	**0.216** (0.011)	**0.213** (0.009)	**0.224** (0.010)
	$d = 5$			**0.245** (0.014)	**0.284** (0.021)	**0.316** (0.025)
FCSE	$d = 1$	**0.207** (0.010)	**0.206** (0.010)	**0.190** (0.008)	**0.214** (0.010)	**0.230** (0.012)
	$d = 2$	**0.215** (0.010)	**0.222** (0.009)	**0.202** (0.009)	**0.197** (0.010)	**0.195** (0.010)
	$d = 3$		**0.216** (0.011)	**0.209** (0.010)	**0.214** (0.010)	**0.207** (0.011)
	$d = 4$		**0.190** (0.007)	**0.223** (0.010)	**0.220** (0.012)	**0.207** (0.010)
	$d = 5$			**0.254** (0.012)	**0.255** (0.015)	**0.302** (0.039)
	ρ	0.044	0.101	0.159	0.216	0.274
RFSIR	$d = 1$	**0.236** (0.011)	**0.222** (0.012)	**0.244** (0.012)	**0.219** (0.010)	**0.221** (0.011)
	$d = 2$	**0.206** (0.011)	**0.224** (0.011)	**0.230** (0.011)	**0.235** (0.011)	**0.236** (0.012)
	$d = 3$	**0.219** (0.009)	**0.218** (0.011)	**0.212** (0.010)	**0.216** (0.009)	**0.232** (0.011)
	$d = 4$	**0.198** (0.010)	**0.215** (0.010)	**0.207** (0.010)	**0.197** (0.008)	**0.189** (0.008)
	$d = 5$	**0.208** (0.010)	**0.193** (0.009)	**0.211** (0.012)	**0.211** (0.011)	**0.234** (0.011)

Table 1: The mean (standard error) of the out-of-sample MSE for predicting logarithm of daily average bike rental counts using projected predictors after different dimension reduction methods.

Remark 1. In dealing with real data, a crucial question is how to select the optimal m.

The selection method provided in Theorem 2 is based on asymptotic theory, which aims to provide minimax optimality results of FSIR under general conditions and is not directly applicable to real data. To date, the problem of selecting the optimal m for a particular data set remains unresolved, as shown in Hall and Horowitz (2007) and Lian (2015).

To utilize the asymptotic results of Theorem 2 for selecting m in practice, we first estimate parameters α and β according to Assumption 4. Specifically, we first obtain the d eigenfunctions, $\widehat{\boldsymbol{\beta}}_k'$ $(k = 1, \ldots, d)$, of $\widehat{\Gamma}_e$ and set $\widehat{\eta}_k = \widehat{\Gamma}^{-1}\widehat{\boldsymbol{\beta}}_k'$. Then we estimate α and β according to $\widehat{\lambda}_j - \widehat{\lambda}_{j+1} \geqslant c_2 j^{-\widehat{\alpha}-1}$ and $|\widehat{b}_{ij}| := \langle \widehat{\eta}_i, \widehat{\phi}_j \rangle \leqslant c_2' j^{-\widehat{\beta}}$ where $\widehat{\Gamma}_e, \widehat{\Gamma}, \widehat{\lambda}_j$ and $\widehat{\phi}_j$ are defined in Algorithm 1. For example, we calculate $\widehat{\alpha} = -(\frac{\ln(\widehat{\lambda}_j - \widehat{\lambda}_{j+1})}{\ln j} + 1)$ and $\widehat{\beta} = -\frac{\ln|\widehat{b}_{ij}|}{\ln j}$ for sufficiently large j respectively. After we get $\widehat{\alpha}$ and $\widehat{\beta}$, we choose m in the interval $[n^{\frac{1}{\widehat{\alpha}+2\widehat{\beta}}}/\log(n), n^{\frac{1}{\widehat{\alpha}+2\widehat{\beta}}} \cdot \log(n)]$ and choose ρ in $[n^{-\frac{\widehat{\alpha}}{\widehat{\alpha}+2\widehat{\beta}}}/\log(n), n^{-\frac{\widehat{\alpha}}{\widehat{\alpha}+2\widehat{\beta}}} \cdot \log(n)]$ (see Lian (2015)). This approach significantly narrows down the choice range for m and ρ and is also consistent with our asymptotic results. In our experiments, feasible values for m were within $\{1, 2, \ldots, 11\}$ and that for ρ were within $[0.015, 0.302]$. For the ease of presentation, we selected 5 representative values each for m and ρ. Detailed results are presented in Table 1.

5. Discussion

In this paper, we established the minimax rate-optimality of FSIR-OT for estimating the functional central space. Specifically, we first prove an upper bound on the convergence rate of the FSIR-OT estimator of the functional central space under very mild assumptions. Then we establish a minimax lower bound on the estimation of the functional central space over a large class of distributions. These two results together show optimality of FSIR-OT.

Our results not only enrich the theoretical understanding of FSIR-OT but also indicate the possibility of extending the findings of multivariate SDR methods to functional data.

There are some open questions related to the findings in this paper. First, the structural dimension d is assumed to be bounded in the current paper. It is still unclear whether this restriction can be relaxed so that the minimax convergence rate of the functional central space estimation can be determined even when d is large (i.e., there is no constant upper bound on d). Second, recent studies have revealed the dependence of the estimation error on the *gSNR* defined as $\lambda_d(\mathrm{Cov}\,(\mathbb{E}[\boldsymbol{X} \mid Y]))$ for multivariate SIR (Lin et al., 2021; Huang et al., 2023)). Exploring the role of gSNR in the estimation of the functional central space will be an interesting next step.

Supplementary Materials

Supplement to "On the Optimality of Functional Sliced Inverse Regression". The supplementary material includes the proofs for all the theoretical results in the paper.

Acknowledgements

Huang's research was supported in part by NUS Start-up Grant A-0004824-00-0 and Singapore Ministry of Education AcRF Tier 1 Grant A-8000466-00-00. Lin's research was supported in part by the National Natural Science Foundation of China (Grant 11971257), Beijing Natural Science Foundation (Grant Z190001), National Key R&D Program of China (2020AAA0105200), and Beijing Academy of Artificial Intelligence. Liu's research was supported in part by the National Science Foundation of the United States (DMS-2015411).

References

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association 91*(435), 983–992.

Cook, R. D. and S. Weisberg (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association 86*(414), 328–332.

Ferré, L. and A.-F. Yao (2003). Functional sliced inverse regression analysis. *Statistics 37*(6), 475–488.

Forzani, L. and R. D. Cook (2007). A note on smoothed functional inverse regression. *Statistica Sinica*, 1677–1681.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91.

Hall, P. and K.-C. Li (1993). On almost linearity of low dimensional projections from high dimensional data. *The annals of Statistics*, 867–889.

Hsing, T. and R. J. Carroll (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 1040–1061.

Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, Volume 997. John Wiley & Sons.

Huang, D., S. Tian, and Q. Lin (2023). Sliced inverse regression with large structural dimensions.

Lei, J. (2014). Adaptive global testing for functional linear models. *Journal of the American Statistical Association 109*(506), 624–634.

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

Li, Y. and T. Hsing (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *The Annals of Statistics 38*(5), 3028–3062.

Lian, H. (2015). Functional sufficient dimension reduction: Convergence rates and multiple functional case. *Journal of Statistical Planning and Inference 167*, 58–68.

Lian, H. and G. Li (2014). Series expansion for functional sufficient dimension reduction. *Journal of Multivariate Analysis 124*, 150–165.

Lin, Q., X. Li, D. Huang, and J. S. Liu (2021). On the optimality of sliced inverse regression in high dimensions. *The Annals of Statistics 49*(1), 1–20.

Lin, Q., Z. Zhao, and J. S. Liu (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics 46*(2), 580–610.

Tan, K., L. Shi, and Z. Yu (2020). Sparse sir: Optimal rates and adaptive estimation. *The Annals of Statistics 48*(1), 64–85.

Wang, G. and H. Lian (2020). Functional sliced inverse regression in a reproducing kernel hilbert space. *Statistica Sinica 30*(1), 17–33.

Wu, Y. and L. Li (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statistica Sinica 2011*(21), 707.

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2009). An adaptive estimation of dimension reduction space. pp. 299–346.

Yang, S. S. (1977). General distribution theory of the concomitants of order statistics. *The Annals of Statistics 5*(5), 996–1002.

Yao, F., E. Lei, and Y. Wu (2015). Effective dimension reduction for sparse functional data. *Biometrika 102*(2), 421–437.

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer.

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

Zhu, L.-X. and K. W. Ng (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, 727–736.

Rui Chen, Center for Statistical Science, Department of Industrial Engineering, Tsinghua University

E-mail: chenrui_fzu@163.com

Songtao Tian, Department of Mathematical Sciences, Tsinghua University

E-mail: tst20@mails.tsinghua.edu.cn

Dongming Huang, Department of Statistics and Data Science, National University of Singapore

E-mail: stahd@nus.edu.sg

Qian Lin, Department of Statistics and Data Science, Tsinghua University; Beijing Academy of Artificial Intelligence, Beijing, 100084, China

E-mail: qianlin@tsinghua.edu.cn

Jun S. Liu, Department of Statistics, Harvard University

E-mail: jliu@stat.harvard.edu