Electronic Journal of Statistics
Vol. 16 (2022) 1997–2032
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS1999

Minimal σ-field for flexible sufficient dimension reduction*

Hanmin Guo and Lin Hou

Center for Statistical Science, Tsinghua University, Beijing, 100084, China
e-mail: ghm17@mails.tsinghua.edu.cn; houl@tsinghua.edu.cn

Yu Zhu

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
e-mail: yuzhu@purdue.edu

Abstract: Sufficient Dimension Reduction (SDR) becomes an important tool for mitigating the curse of dimensionality in high dimensional regression analysis. Recently, Flexible SDR (FSDR) has been proposed to extend SDR by finding lower dimensional projections of *transformed* explanatory variables. The dimensions of the projections however cannot fully represent the extent of data reduction FSDR can achieve. As a consequence, optimality and other theoretical properties of FSDR are currently not well understood. In this article, we propose to use the σ-field associated with the projections, together with their dimensions to fully characterize FSDR, and refer to the σ-field as the FSDR σ-field. We further introduce the concept of minimal FSDR σ-field and consider FSDR projections with the minimal σ-field optimal. Under some mild conditions, we show that the minimal FSDR σ-field exists, attaining the lowest dimensionality at the same time. To estimate the minimal FSDR σ-field, we propose a two-stage procedure called the Generalized Kernel Dimension Reduction (GKDR) method and partially establish its consistency property under weak conditions. Extensive simulation experiments demonstrate that the GKDR method can effectively find the minimal FSDR σ-field and outperform other existing methods. The application of GKDR to a real life air pollution data set sheds new light on the connections between atmospheric conditions and air quality.

MSC2020 subject classifications: Primary 62B05; secondary 62J02.
Keywords and phrases: High dimensional regression analysis, univariate transformation, sufficient predictor, Reproducing Kernel Hilbert Space, conditional entropy.

Received June 2021.

Contents

*Part of Y.Z.'s work was conducted during his visits to the Center for Statistical Science, Tsinghua University, and the School of Sciences, Huzhou Normal University. L.H. acknowledges research support from the National Natural Science Foundation of China (Grant No. 12071243).

1. Introduction

Statistical analysis of high dimensional data is known to suffer from the curse of dimensionality, and dimension reduction methods are usually used to reduce the dimensionality so as to mitigate the curse. Li [17] proposed *Sufficient Dimension Reduction (SDR)* as an effective approach for reducing dimensionality in high dimensional regression analysis. Let Y be the response, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ the p-dimensional vector of explanatory variables, and $B \in \mathbb{R}^{p\times d}$ a matrix of p rows and d columns. If given $B^T\boldsymbol{X}$, Y and $\boldsymbol{X}$ are independent with each other, that is,

$$Y \perp\!\!\!\perp \boldsymbol{X} | B^T\boldsymbol{X}, \tag{1}$$

where $\perp\!\!\!\perp$ denotes "independent", then the column space of B, denoted by $span(B)$, is called an *SDR subspace*. SDR subspaces are not unique. Under some mild conditions, the intersection of all SDR subspaces remains an SDR subspace, which is called the *central subspace* and denoted by $S_{Y|\boldsymbol{X}}$ [4]. After $S_{Y|\boldsymbol{X}}$ is obtained, subsequent analyses can be applied to $P_{S_{Y|\boldsymbol{X}}}\boldsymbol{X}$, which is the projection of $\boldsymbol{X}$ onto the central subspace $S_{Y|\boldsymbol{X}}$.

Another approach to coping with the curse of dimensionality in high dimensional regression analysis is to impose the assumption that the dependence of Y on $\boldsymbol{X}$ admits a lower dimensional configuration. The ACE algorithm [2] and the

generalized additive model [12] are two methods following this approach. Both of the methods postulate the following additive regression model

$$h(Y) = \alpha + \sum_{i=1}^{p} \phi_i(X_i) + \epsilon,$$

where $h, \phi_1, \ldots, \phi_p$ are unknown univariate functions, α is the intercept, and ϵ is the error term. The additive regression model can be considered an extension of covariate transformations commonly used in linear regression analysis. It does not suffer from the curse of dimensionality and is more flexible in practice.

The idea of using transformed explanatory variables for linear regression can also be employed for sufficient dimension reduction. Wang and Zhu [23] proposed the *flexible SDR* model given below.

$$Y \perp\!\!\!\perp \boldsymbol{X} | B^T \boldsymbol{\phi}(\boldsymbol{X}), \tag{2}$$

where B is a $p \times d$ matrix as before and $\boldsymbol{\phi}(\boldsymbol{X}) = \big(\phi_1(X_1), \cdots, \phi_p(X_p)\big)^T$. Here we refer to $B^T\boldsymbol{\phi}(\boldsymbol{X})$ satisfying (2) as a *sufficient predictor vector*. Compared with the original SDR, the flexible SDR model can provide more flexibility in modeling the relationship between Y and $\boldsymbol{X}$, and lead to further dimensionality reduction. For example, consider the model $Y = X_1^2/\big(\sin(X_2) + X_3^2 + 2\big) + \epsilon$, where $\boldsymbol{X} = (X_1, \ldots, X_{10})^T$, and $\epsilon \perp\!\!\!\perp \boldsymbol{X}$. Under the original SDR, it can be shown that the central subspace is the three-dimensional subspace spanned by $\boldsymbol{e}_1, \boldsymbol{e}_2$ and $\boldsymbol{e}_3$, where $\boldsymbol{e}_i$ is the column vector with the i-th entry being 1 and the others being 0 for $i = 1, 2, 3$. The projection of $\boldsymbol{X}$ onto the central subspace yields a three-dimensional space, which is the sample space for $(X_1, X_2, X_3)^T$; Under the flexible SDR, however, we can set $\boldsymbol{\phi}(\boldsymbol{X}) = \big(X_1, \sin(X_2), X_3^2, X_4, \ldots, X_{10}\big)^T$ and $B = (e_1, e_2 + e_3)$, and then $Y \perp\!\!\!\perp \boldsymbol{X} | B^T\boldsymbol{\phi}(\boldsymbol{X})$. $B^T\boldsymbol{\phi}(\boldsymbol{X})$ can be considered a two-dimensional sufficient predictor vector, and we can rely on $B^T\boldsymbol{\phi}(\boldsymbol{X})$ to explore the relationship between Y and $\boldsymbol{X}$ instead of the original $\boldsymbol{X}$. Therefore, the dimensionality can be reduced to two after flexible SDR is performed.

For the example discussed above, we can find another set of transformations $\tilde{\boldsymbol{\phi}}(\boldsymbol{X}) = \big(X_1^2, \sin(X_2), X_3^2, X_4, \ldots, X_{10}\big)^T$, which also satisfies $Y \perp\!\!\!\perp \boldsymbol{X} | B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$. Therefore, both $B^T\boldsymbol{\phi}(\boldsymbol{X})$ and $B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$ are sufficient predictor vectors. Wang and Zhu [23] used the dimensions of B to characterize flexible SDR. In this example, the dimensions of B cannot be used to discriminate between $B^T\boldsymbol{\phi}(\boldsymbol{X})$ and $B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$ because they share exactly the same B. Notice the only difference between $B^T\boldsymbol{\phi}(\boldsymbol{X})$ and $B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$ is that $\phi_1(X_1) = X_1$ whereas $\tilde{\phi}_1(X_1) = X_1^2$. From the model, Y depends on X_1 only through X_1^2, and the sign of X_1 is not important. Therefore, $B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$ should be considered a further reduction from $B^T\boldsymbol{\phi}(\boldsymbol{X})$. The reduction from $B^T\boldsymbol{\phi}(\boldsymbol{X})$ to $B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$ can in fact be best characterized by their respective σ-fields. Let $\sigma_{B,\boldsymbol{\phi}}$ and $\sigma_{B,\tilde{\boldsymbol{\phi}}}$ denote the σ-fields associated with $B^T\boldsymbol{\phi}(\boldsymbol{X})$ and $B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$, respectively. It can be shown that $\sigma_{B,\tilde{\phi}}$ is a proper sub-field of $\sigma_{B,\phi}$, that is, $\sigma_{B,\tilde{\phi}} \subset \sigma_{B,\phi}$. The size of a σ-field can reflect the amount of information the σ-field contains, and the smaller the σ-field is,

the more concentrated the information it contains [1]. From the perspective of dimension reduction, therefore, $B^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$ should be preferred over $B^T\boldsymbol{\phi}(\boldsymbol{X})$.

The use of σ-fields for facilitating SDR was first introduced by Lee et al. [16]. Let $\sigma(\boldsymbol{X})$ denote the σ-field for $\boldsymbol{X}$, and $\mathcal{G}$ any sub σ-field of $\sigma(\boldsymbol{X})$. If

$$Y \perp\!\!\!\perp \boldsymbol{X}|\mathcal{G}, \tag{3}$$

then $\mathcal{G}$ is called an *SDR σ-field* for the dependence of Y on $\boldsymbol{X}$. Under some mild conditions, the intersection of all SDR σ-fields still satisfies the conditional independence assertion (3), and is called the *central σ-field* [16]. SDR σ-fields retain sufficient information stored in $\boldsymbol{X}$ for predicting Y, and the central σ-field is the smallest SDR σ-field. The central σ-field has the advantage of achieving the largest possible data reduction. This advantage however comes with a trade-off, which is that it no longer admits the concept of dimensionality. In the example discussed above, it can be shown that the central σ-field is the σ-field associated with $X_1^2/\big(\sin(X_2)+X_3^2+2\big)$, and it is not clear how to properly define the dimensionality of this σ-field. Furthermore, finding the central σ-field is equivalent to directly performing nonparametric regression, and it does not produce a set of sufficient predictor variables as in the original SDR for subsequent data exploration and analysis such as visualization and model construction.

In this article, instead of using the dimensions of B, we propose to use the σ-field associated with $B^T\boldsymbol{\phi}(\boldsymbol{X})$ to characterize flexible SDR, which can be equivalently redefined as follows.

$$Y \perp\!\!\!\perp \boldsymbol{X}|\sigma_{B,\boldsymbol{\phi}}, \tag{4}$$

where $\sigma_{B,\boldsymbol{\phi}}$ is the σ-field associated with $B^T\boldsymbol{\phi}(\boldsymbol{X})$. We refer to $\sigma_{B,\boldsymbol{\phi}}$ as the Flexible SDR σ-field or FSDR σ-field, and call $(B,\boldsymbol{\phi})$ a *generating pair* of $\sigma_{B,\boldsymbol{\phi}}$. An FSDR σ-field can be considered a compromise between the original SDR [17] and the general SDR σ-field [16]. Unlike the general SDR σ-field, an FSDR σ-field is equipped with a generating pair B and $\boldsymbol{\phi}$, and thus preserves the concept of dimensionality and a set of sufficient predictor variables (i.e., $B^T\boldsymbol{\phi}$). For convenience, we refer to the column rank of B as the dimension of the FSDR σ-field $\sigma_{B,\boldsymbol{\phi}}$.

FSDR σ-fields are not unique, neither are the generating pairs of a given FSDR σ-field. Different FSDR σ-fields represent different degrees of data reduction. Intuitively, the smallest FSDR σ-field should be considered optimal, because it achieves the largest data reduction under the framework of flexible SDR. We refer to the smallest FSDR σ-field as the *minimal FSDR σ-field*. The existence of the minimal FSDR σ-field is not obvious. In a properly defined class of FSDR σ-fields, we will show that the minimal FSDR σ-field exists, and it also has the smallest dimension; See Section 2. Therefore, the minimal FSDR σ-field should be the inference target under the flexible SDR.

In the literature, a variety of methods have been proposed to perform estimation for SDR, including sliced inverse regression (SIR; [17]), sliced average variance estimation (SAVE; [5]), and many others. These methods mainly take the inverse regression approach and only utilize the first couple of moments

of the conditional distribution of $\boldsymbol{X}$ given Y. They can consistently estimate the central space only under various conditions such as the linearity condition. There are some other estimation methods for SDR, which do not follow the inverse regression. The Fourier method [29] and the Minimum Average Variance Estimation (MAVE) method [27] are two such methods. Also, there are some recent works that introduce kernel methods into sufficient dimension reduction, such as the Kernel Sliced Inverse Regression (KSIR) method [26] and its extension called the Kernel Additive Sliced Inverse Regression (KASIR) method [18], the Kernel Dimension Reduction (KDR) method [10], and the Gradient Based Kernel Dimension Reduction method [11]. Specifically, the KDR method characterizes the conditional independence assertion (1) through the conditional covariance operators between the Reproducing Kernel Hilbert Spaces (RKHS) of Y and $\boldsymbol{X}$, respectively. The key advantage of the KDR method is that it no longer requires the linearity condition, and theoretically it can fully recover the central subspace when the dimension of the true central subspace is known. Because of this advantage, we extend the KDR method to perform inference for the minimal FSDR σ-field. However, it should be noted that the inference for the minimal FSDR σ-field is not restricted to the KDR method and can be also based on other methods.

In this article, we prove that under some mild conditions, the minimal FSDR σ-field uniquely exists and has the smallest dimension. This result is placed in Section 2. We propose the *Generalized Kernel Dimension Regression* (GKDR) method, an extension of the KDR method, to estimate the minimal FSDR σ-field in Section 3. The GKDR method is a two-stage approach. In the first stage, it estimates multiple FSDR σ-fields with the same dimension as the minimal FSDR σ-field through the minimization of conditional covariance operator between RKHSs. To perform nonparametric estimation, B-spline functions are utilized. We prove that under some conditions, the GKDR Stage I estimator has some good consistency properties. In the second stage, the GKDR method estimates the minimal FSDR σ-field through the maximization of conditional entropy among all the FSDR σ-fields obtained in the first-stage. In Section 4, we discuss some implementation details and provide numerical results for a number of simulation studies and a real data application of the proposed method. We conclude the article with some future research directions in Section 5. All the technical proofs of the theorems and propositions are presented in the Appendix A.

2. Existence of the minimal FSDR σ-field

Let $\mathcal{X}$ and $\mathcal{Y}$ be the supports of $\boldsymbol{X}$ and Y, and $P_{\boldsymbol{X}}$ and P_Y the probability distributions of $\boldsymbol{X}$ and Y, respectively. Let $P_{\boldsymbol{X}Y}$ be the joint probability distribution of $\boldsymbol{X}$ and Y. Suppose $\mathcal{F}_{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{Y}}$ are the Borel σ-fields over $\mathcal{X}$ and $\mathcal{Y}$, respectively. For $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, $P_{Y|\boldsymbol{X}}(\cdot|\boldsymbol{x})$ and $P_{\boldsymbol{X}|Y}(\cdot|y)$ denote the conditional probability distributions of Y given $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{X}$ given $Y = y$, respectively. We first state two required assumptions for the existence of the minimal FSDR σ-field.

Assumption 1. *The family of probability distributions* $\{P_{\boldsymbol{X}|Y}(\cdot|y) : y \in \mathcal{Y}\}$ *is dominated by a* σ*-finite measure.*

Assumption 1 is adopted from Lee et al. [16]. Under this assumption, for any two SDR σ-fields $\mathcal{G}_1, \mathcal{G}_2 \subseteq \sigma(\boldsymbol{X})$, their intersection remains an SDR σ-field. Let $\boldsymbol{X}_{(-i)}$ denote the sub-vector of $\boldsymbol{X}$ excluding the i-th entry X_i, and $\boldsymbol{\phi}_{(-i)}(\boldsymbol{X}_{(-i)})$ the sub-vector of $\boldsymbol{\phi}(\boldsymbol{X})$ excluding the i-th entry $\phi_i(X_i)$. Let $\mathcal{X}_{(-i|x_i)}$ denote the support of the conditional distribution of $\boldsymbol{X}_{(-i)}$ given $X_i = x_i$.

Assumption 2. *If* $\mathcal{X}_{(-i|x_i)} \neq \varnothing$ *and* $\mathcal{X}_{(-i|\tilde{x}_i)} \neq \varnothing$*, then* $\mathcal{X}_{(-i|x_i)} \cap \mathcal{X}_{(-i|\tilde{x}_i)} \neq \varnothing$.

If the support of a random vector is the whole Euclidean space $\mathbb{R}^p$, sphere, hemisphere, or cube, then this random vector satisfies Assumption 2. In the following proposition, we state that if a random vector $\boldsymbol{X}$ satisfies Assumption 2, then the components of $\boldsymbol{X}$ are not nonlinearly confounded with each other, that is, none of them is a function of some other components.

Proposition 1. *Suppose Assumption 2 holds. For any* i *with* $1 \leq i \leq p$*, if there are two transformations* ζ_i *and* ψ_i *satisfying*

$$E[\zeta_i(X_i)] = 0,\ E[\psi_i(\boldsymbol{X}_{(-i)})] = 0,\ \textit{and}\ \zeta_i(X_i) + \psi_i(\boldsymbol{X}_{(-i)}) = 0\ \textit{almost surely},$$

then both ζ_i *and* ψ_i *are zero almost surely.*

The proof of Proposition 1 can be found in Appendix A.1. Note that the proofs of the remaining theorems throughout this article are also placed in the Appendix A. Recall that $\sigma_{B,\boldsymbol{\phi}}$ is the σ-field associated with $B^T\phi(\boldsymbol{X})$. In the rest of this article, we sometimes write it as $\sigma\big(B^T\phi(\boldsymbol{X})\big)$ for convenience. The matrix B plays an important role in the way that X_i's affect $\sigma_{B,\boldsymbol{\phi}}$. We next introduce a partition of the indices $\{1, 2, \ldots, p\}$ based on B. We define $\mathcal{I}_B = \{i : \boldsymbol{e}_i \in span(B)\}$ and $\mathcal{K}_B = \{k : B^T\boldsymbol{e}_k = 0\}$, and let $\mathcal{J}_B = \{1, 2, \ldots, p\}\backslash(\mathcal{I}_B \cup \mathcal{K}_B) = \{j : B^T\boldsymbol{e}_j \neq 0\}\backslash\mathcal{I}_B$. It can be verified that the partition is unique with respect to the column space of B, that is, for any two matrices B and B', if $span(B) = span(B')$, then $\mathcal{I}_B = \mathcal{I}_{B'}, \mathcal{J}_B = \mathcal{J}_{B'}$, and $\mathcal{K}_B = \mathcal{K}_{B'}$.

In general, the partition of $\mathcal{I}_B$, $\mathcal{J}_B$, and $\mathcal{K}_B$ determines how the σ-fields of $\phi(X_i)$'s are related to $\sigma_{B,\boldsymbol{\phi}}$. Consider the example discussed in Section 1: $Y = X_1^2/\big(\sin(X_2) + X_3^2 + 2\big) + \epsilon$. An FSDR σ-field is $\sigma_{B,\boldsymbol{\phi}} = \sigma\big(\phi_1(X_1), \phi_2(X_2) + \phi_3(X_3)\big)$, where $B^T = \begin{pmatrix} 1,0,0,0 \\ 0,1,1,0 \end{pmatrix}$, $\phi_1(X_1) = X_1$, $\phi_2(X_2) = \sin(X_2)$, $\phi_3(X_3) = X_3^2$, and $\phi_4(X_4) = 0$. For this FSDR σ-field, the partition is $\mathcal{I}_B = \{1\}$, $\mathcal{J}_B = \{2, 3\}$, and $\mathcal{K}_B = \{4\}$. Correspondingly, X_1, X_2, X_3, and X_4 are partitioned into $\{X_1\}$, $\{X_2, X_3\}$, and $\{X_4\}$. It can be shown that the σ-field of $\phi_1(X_1)$ denoted as $\sigma(\phi_1(X_1))$ is a sub-σ-field of $\sigma_{B,\boldsymbol{\phi}}$, that is, $\sigma\big(\phi_1(X_1)\big) \subseteq \sigma_{B,\boldsymbol{\phi}}$, and $\sigma\big(\phi_2(X_2) + \phi_3(X_3)\big) \subseteq \sigma_{B,\boldsymbol{\phi}}$; However, $\sigma(\phi_2(X_2)) \nsubseteq \sigma_{B,\boldsymbol{\phi}}$, $\sigma\big(\phi_3(X_3)\big) \nsubseteq \sigma_{B,\boldsymbol{\phi}}$, and $\phi_4(X_4)$ is irrelevant to $\sigma_{B,\boldsymbol{\phi}}$. This property holds in general cases. The σ-fields of $\phi_i(X_i)$'s with $i \in \mathcal{I}_B$ are contained in $\sigma_{B,\boldsymbol{\phi}}$; The σ-fields of the linear combinations of $\phi_j(X_j)$ with $j \in \mathcal{J}_B$, but not the individual σ-fields of $\phi_j(X_j)$'s, are contained in $\sigma_{B,\boldsymbol{\phi}}$; And the σ-fields of $\phi_k(X_k)$'s with $k \in \mathcal{K}_B$ are irrelevant to $\sigma_{B,\boldsymbol{\phi}}$.

Suppose the cardinalities of $\mathcal{I}_B$, $\mathcal{J}_B$, and $\mathcal{K}_B$ are p_1, p_2, and p_3, respectively. Without loss of generality, we assume that the indices in $\mathcal{I}_B$ are smaller than those in $\mathcal{J}_B$, and the indices in $\mathcal{J}_B$ are smaller than those in $\mathcal{K}_B$. This is equivalent to the assumption that B has the following standardized form $\begin{pmatrix} I_{p_1}, & 0 \\ 0, & B_{\mathcal{J}} \\ 0, & 0 \end{pmatrix}$, where I_{p_1} is the $p_1 \times p_1$ identity matrix, $B_{\mathcal{J}}$ is a $p_2 \times (d-p_1)$ matrix, and for each column of $B_{\mathcal{J}}$ there exist at least two nonzero entries. Let $\boldsymbol{X}_{\mathcal{I}_B}$ denote the subvector of $\boldsymbol{X}$ indexed by $\mathcal{I}_B$, and $\phi_{\mathcal{I}_B}(\boldsymbol{X}_{\mathcal{I}_B})$ the subvector of $\boldsymbol{\phi}(\boldsymbol{X})$ indexed by $\mathcal{I}_B$. The same notational rules apply to $\boldsymbol{X}_{\mathcal{J}_B}$, $\boldsymbol{X}_{\mathcal{K}_B}$, $\phi_{\mathcal{J}_B}(\boldsymbol{X}_{\mathcal{J}_B})$, and $\phi_{\mathcal{K}_B}(\boldsymbol{X}_{\mathcal{K}_B})$. Then we have $\sigma_{B,\boldsymbol{\phi}} = \sigma\big(\bigcup_{i\in\mathcal{I}_B} \phi_i(X_i), B_{\mathcal{J}}^T \boldsymbol{\phi}_{\mathcal{J}_B}(\boldsymbol{X}_{\mathcal{J}_B})\big)$. We next define a proper class of FSDR σ-fields. Let

$$\mathcal{A} = \big\{\sigma_{B,\boldsymbol{\phi}} : \ Y \perp\!\!\!\perp X | \sigma_{B,\boldsymbol{\phi}}, \text{ and the support of } \phi_j(X_j) \text{ is an interval for } j \in \mathcal{J}_B\big\}.$$

The class $\mathcal{A}$ is rich enough to contain most FSDR σ-fields. For example, if the transformations $\phi_1(X_1), \ldots, \phi_p(X_p)$ are continuous and the conditional independence assertion (2) is satisfied, then $\sigma_{B,\boldsymbol{\phi}}$ belongs to $\mathcal{A}$. The purpose of requiring that the supports of $\phi_j(X_j)'s$ are intervals is to exclude degenerate cases involving both continuous and discrete transformations. Let's consider a counterexample. Suppose

$$Y = \phi_1(X_1) + \phi_2(X_2) + \epsilon,$$

where $\phi_1(X_1) = e^{X_1}/(e^{X_1}+1)$, $\phi_2(X_2) = I(X_2 > 0)$, and $I(\cdot)$ is the indicator function. Here $\phi_2(X_2)$ only takes value in $\{0, 1\}$, so the support of $\phi_2(X_2)$ is not an interval. Let B_1 be the 2×2 identity matrix I_2, and $B_2^T = (1, 1)$. It can be verified that both $\sigma_{B_1,\boldsymbol{\phi}}$ and $\sigma_{B_2,\boldsymbol{\phi}}$ are FSDR σ-fields and they are identical. Note that the column rank of B_1 is two and the column rank of B_2 is one. Although the rank of B_1 is higher than that of B_2, the two FSDR σ-fields are the same. In general, when discrete transformations are involved, the dimension can be further reduced while the FSDR σ-field remains the same. In this article, we require that the support of $\phi_j(X_j)$ should be an interval for $j \in \mathcal{J}_B$ to exclude the degenerate cases such as the example discussed above. We now state the existence theorem of the minimal FSDR σ-field as follows.

Theorem 2. *Under Assumptions 1 and 2, there exists a unique minimal FSDR σ-field $\sigma_{B^*,\boldsymbol{\phi}^*}$ in $\mathcal{A}$ such that*

$$\sigma_{B^*,\boldsymbol{\phi}^*} \subseteq \sigma_{B,\boldsymbol{\phi}}, \text{ for any } \sigma_{B,\boldsymbol{\phi}} \in \mathcal{A}. \tag{5}$$

Furthermore, suppose $(\tilde{B}, \tilde{\boldsymbol{\phi}})$ is another generating pair of the minimal FSDR σ-field $\sigma_{B^*,\boldsymbol{\phi}^*}$. Then $span(\tilde{B}) = span(B^*)$, $\sigma\big(\tilde{\phi}_i(X_i)\big) = \sigma\big(\phi_i^*(X_i)\big)$ for $i \in \mathcal{I}_{B^*}$, and there exist constant numbers u_j and v_j such that $\tilde{\phi}_j = u_j * \phi_j^* + v_j$ for $j \in \mathcal{J}_{B^*}$.

Note that FSDR σ-fields in class $\mathcal{A}$ have a partial order instead of a linear order, and Zorn's Lemma is required in the proof of Theorem 2. Theorem 2

implies that although the generating pair $(B^*, \boldsymbol{\phi}^*)'s$ are not unique, the column space $span(B^*)$ is unique. Furthermore, the univariate transformation $\phi_i^* s$ can be determined up to a one-to-one mapping for $i \in \mathcal{I}_{B^*}$, and the univariate transformation $\phi_j^* s$ can be determined up to a linear transformation for $j \in \mathcal{J}_{B^*}$. $\sigma_{B^*,\boldsymbol{\phi}^*}$ in Theorem 2 is extraordinary since it achieves the smallest $\sigma_{B,\boldsymbol{\phi}}$. In addition to $\sigma_{B^*,\boldsymbol{\phi}^*}$, the FSDR σ-fields with the same dimension as $\sigma_{B^*,\boldsymbol{\phi}^*}$ are also of interest and will play a critical role in estimating $\sigma_{B^*,\boldsymbol{\phi}^*}$, as will be shown in Section 3.4. The generating pairs of these FSDR σ-fields possess similar properties as that of $\sigma_{B^*,\boldsymbol{\phi}^*}$, which we state in the following theorem.

Theorem 3. *Suppose $(\tilde{B}, \tilde{\boldsymbol{\phi}})$ is a generating pair of an FSDR σ-field with the same dimension as $\sigma_{B^*,\boldsymbol{\phi}^*}$. Then $span(\tilde{B}) = span(B^*)$, $\sigma\big(\tilde{\phi}_i(X_i)\big) \supseteq \sigma(\phi_i^*(X_i))$ for $i \in \mathcal{I}_{B^*}$, and there exist constant numbers u_j and v_j such that $\tilde{\phi}_j = u_j * \phi_j^* + v_j$ for $j \in \mathcal{J}_{B^*}$.*

Theorem 3 implies that for $\sigma_{B^*,\boldsymbol{\phi}^*}$, not only the dimension is the smallest, but also the univariate transformations are the coarsest, that is, the univariate transformations of $\sigma_{B^*,\boldsymbol{\phi}^*}$ can be represented as functions of those of other FSDR σ-fields. We will utilize this property for estimating $\sigma_{B^*,\boldsymbol{\phi}^*}$, as will be discussed in Sections 3.1 and 3.4.

3. Estimation scheme for $\boldsymbol{\sigma_{B^*,\phi^*}}$

Suppose the flexible dimension reduction model (2) holds, and Assumptions 1 and 2 are satisfied. In Section 2, we have proved that the minimal FSDR σ-field $\sigma_{B^*,\boldsymbol{\phi}^*}$ exists and is unique, and hereafter we always assume that d_0, the dimension of $\sigma_{B^*,\boldsymbol{\phi}^*}$, is already known. In this section, we present the GKDR method for estimating $\sigma_{B^*,\boldsymbol{\phi}^*}$, as discussed in Section 1.

The GKDR method uses a two-stage approach. In the first stage, it estimates the generating pair $(B, \boldsymbol{\phi})'s$ of the FSDR σ-fields with the same dimension as $\sigma_{B^*,\boldsymbol{\phi}^*}$. We present this procedure at the population level in Section 3.1, and then we present its sample version and consistency result in Sections 3.2 and 3.3, respectively. In the second stage, the GKDR method incorporates conditional entropy to estimate one representing generating pair $(B^*, \boldsymbol{\phi}^*)$ of the minimal FSDR σ-field, based on the generating pair $(B, \boldsymbol{\phi})'s$ obtained in the first stage. This procedure is presented in Section 3.4.

3.1. Stage I of GKDR: Population level

For an integer $d \le p$, let $\mathbb{S}_d^p(\mathbb{R})$ be the Stiefel manifold defined as follows.

$$\mathbb{S}_d^p(\mathbb{R}) = \{B \in \mathbb{R}^{p\times d} : B^T B = I_d\}.$$

Let $\tilde{L}^2(P_{X_i})$ and $\tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$ be the families of normalized L^2 functions defined as follows.

$$\tilde{L}^2(P_{X_i}) = \big\{\phi_i :\ E\big[\phi_i(X_i)\big] = 0,\ Var\big[\phi_i(X_i)\big] = 1\big\},\ \text{for } i = 1, \ldots, p;$$

$$\tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}}) = \big\{(\phi_1, \ldots, \phi_p)^T : \phi_i \in \tilde{L}^2(P_{X_i}),\ \text{for } i = 1, \ldots, p\big\}.$$

Let $\mathbb{S}_d^p(\mathbb{R}) \times \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$ be the family of normalized generating pair $(B, \boldsymbol{\phi})'s$. Suppose $(B^*, \boldsymbol{\phi}^*) \in \mathbb{S}_{d_0}^p(\mathbb{R}) \times \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$ is a normalized generating pair of $(B^*, \boldsymbol{\phi}^*)$. We define $\mathbb{B}_{d_0}^p$, which is a subset of $\mathbb{S}_{d_0}^p(\mathbb{R})$, and Φ_0 and Φ_1, which are two subsets of $\tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$, as follows:

$$\mathbb{B}_{d_0}^p = \big\{B \in \mathbb{S}_{d_0}^p(\mathbb{R}) : span(B) = span(B^*)\big\};$$
$$\Phi_0 = \big\{\boldsymbol{\phi} \in \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}}) : \sigma\big(\phi_i(X_i)\big) = \sigma\big(\phi_i^*(X_i)\big) \text{ for } i \in \mathcal{I}_{B^*}, \text{ and } \phi_j = \phi_j^* \text{ for } j \in \mathcal{J}_{B^*}\big\};$$
$$\Phi_1 = \big\{\boldsymbol{\phi} \in \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}}) : \sigma\big(\phi_i(X_i)\big) \supseteq \sigma\big(\phi_i^*(X_i)\big) \text{ for } i \in \mathcal{I}_{B^*}, \text{ and } \phi_j = \phi_j^* \text{ for } j \in \mathcal{J}_{B^*}\big\}.$$

Let $\Theta_0 = \mathbb{B}_{d_0}^p \times \Phi_0$ and $\Theta_1 = \mathbb{B}_{d_0}^p \times \Phi_1$. Note that $\Phi_0 \subseteq \Phi_1$; Therefore we have $\Theta_0 \subseteq \Theta_1$. From Theorem 2, we know that Θ_0 is the family of the normalized generating pairs of $\sigma_{B^*, \boldsymbol{\phi}^*}$. And from Theorem 3, we know that Θ_1 is the family of normalized generating pairs of the FSDR σ-fields which have the same dimension as $\sigma_{B^*, \boldsymbol{\phi}^*}$.

In the rest of this section, we will focus on estimating Θ_1, and the estimation of Θ_0 will be placed in Section 3.4. Let $\mathcal{H}_{\mathcal{X}}$ be the RKHS on $\mathcal{X}$ generated by a positive definite kernel function $k_{\mathcal{X}}$, and $\mathcal{H}_{\mathcal{Y}}$ the RKHS on $\mathcal{Y}$ generated by another positive definite kernel function $k_{\mathcal{Y}}$. The positive definite kernel functions $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are assumed to satisfy the following conditions.

$$E_{\boldsymbol{X}}\big[k_{\mathcal{X}}(\boldsymbol{X}, \boldsymbol{X})\big] < \infty \text{ and } E_Y\big[k_{\mathcal{Y}}(Y, Y)\big] < \infty. \tag{6}$$

Following Fukumizu et al. [10], we define the cross-covariance operator $\Sigma_{Y\boldsymbol{X}} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ as follows.

$$\langle g, \Sigma_{Y\boldsymbol{X}} f\rangle_{\mathcal{H}_{\mathcal{Y}}} = E_{\boldsymbol{X}Y}\Big[\Big(f(\boldsymbol{X}) - E_{\boldsymbol{X}}\big[f(\boldsymbol{X})\big]\Big)\Big(g(Y) - E_Y\big[g(Y)\big]\Big)\Big], \tag{7}$$

for any $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. The variance operators $\Sigma_{\boldsymbol{X}\boldsymbol{X}} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{X}}$ and $\Sigma_{YY} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ can be similarly defined. Further we define the conditional covariance operator $\Sigma_{YY|\boldsymbol{X}} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ as

$$\Sigma_{YY|\boldsymbol{X}} = \Sigma_{YY} - \Sigma_{Y\boldsymbol{X}}\Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1}\Sigma_{\boldsymbol{X}Y}. \tag{8}$$

Note that $\Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1}$ is an abuse of notation because in general $\Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1}$ may not exist. The formal definition can be found in [10], and we omit it here for conciseness.

Let $k_d(\cdot, \cdot)$ be a positive definite kernel function on $\mathbb{R}^d$, such that for any $B \in \mathbb{S}_d^p(\mathbb{R})$ and $\boldsymbol{\phi} \in \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$, the following condition holds.

$$E_{\boldsymbol{X}}\Big[k_d\big(B^T\boldsymbol{\phi}(\boldsymbol{X}), B^T\boldsymbol{\phi}(\boldsymbol{X})\big)\Big] < \infty. \tag{9}$$

Let $\mathcal{H}_{k_d}$ be the RKHS on $\mathbb{R}^d$ generated by the kernel function k_d. For a given normalized generating pair $(B, \boldsymbol{\phi}) \in \mathbb{S}_d^p(\mathbb{R}) \times \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$, we define $k_{\mathcal{X}}^{B,\boldsymbol{\phi}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = k_d\big(B^T\boldsymbol{\phi}(\boldsymbol{x}),\ B^T\boldsymbol{\phi}(\tilde{\boldsymbol{x}})\big)$. It can be verified that $k_{\mathcal{X}}^{B,\boldsymbol{\phi}}$ is a positive definite kernel

function on $\mathcal{X}$. Let $\mathcal{H}_{\mathcal{X}}^{B,\phi}$ be the RKHS on $\mathcal{X}$ generated by the kernel function $k_{\mathcal{X}}^{B,\phi}$. It can be shown that for any $f \in \mathcal{H}_{\mathcal{X}}^{B,\phi}$, one can always find a function $g \in \mathcal{H}_{k_d}$ such that $f(\boldsymbol{x}) = g\big(B^T\boldsymbol{\phi}(\boldsymbol{x})\big)$ for any $\boldsymbol{x} \in \mathcal{X}$. The cross-covariance operator $\Sigma_{Y\boldsymbol{X}}^{B,\boldsymbol{\phi}} : \mathcal{H}_{\mathcal{X}}^{B,\phi} \to \mathcal{H}_{\mathcal{Y}}$ and the variance operator $\Sigma_{\boldsymbol{X}\boldsymbol{X}}^{B,\boldsymbol{\phi}} : \mathcal{H}_{\mathcal{X}}^{B,\phi} \to \mathcal{H}_{\mathcal{X}}^{B,\phi}$ can be similarly defined as $\Sigma_{Y\boldsymbol{X}}$ and $\Sigma_{\boldsymbol{X}\boldsymbol{X}}$, respectively. We further define the conditional covariance operator $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ as

$$\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}} = \Sigma_{YY} - \Sigma_{Y\boldsymbol{X}}^{B,\boldsymbol{\phi}}[\Sigma_{\boldsymbol{X}\boldsymbol{X}}^{B,\boldsymbol{\phi}}]^{-1}\Sigma_{\boldsymbol{X}Y}^{B,\boldsymbol{\phi}}. \tag{10}$$

The conditional covariance operator $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$ is defined for any normalized generating pair $(B,\boldsymbol{\phi}) \in \mathbb{S}_d^p(\mathbb{R}) \times \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$. Recall that Θ_1 is the family of normalized generating pair $(B,\boldsymbol{\phi})'s$ of the FSDR σ-fields with dimension d_0. In fact, Θ_1 can be characterized by the solution set of a minimization problem of $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$, which will be presented in Theorem 4 below. Before that, we require an assumption that is equivalent to Assumption (A-2) proposed in [10]. This assumption guarantees that the RKHS is rich enough to approximate any L^2 function. Let $\mathcal{H}+\mathbb{R}$ denote the direct sum of the RKHS $\mathcal{H}$ and the real number space $\mathbb{R}$, and $P_{B,\boldsymbol{\phi}}$ the probability distribution on $\mathcal{X}$ induced from the projection $BB^T\boldsymbol{\phi} : \mathcal{X} \to \mathcal{X}$.

Assumption 3. *$\mathcal{H}_{\mathcal{X}}+\mathbb{R}$ is dense in $L^2(P_{\boldsymbol{X}})$, and $\mathcal{H}_{\mathcal{X}}^{B,\boldsymbol{\phi}}+\mathbb{R}$ is dense in $L^2(P_{B,\boldsymbol{\phi}})$ for every $B \in \mathbb{S}_d^p(\mathbb{R})$ and $\boldsymbol{\phi} \in \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$.*

We need to introduce one more concept. Let $(\Omega,\mathcal{F})$ be a measurable space, and $\mathcal{H}$ a RKHS on Ω generated by a bounded kernel function k. $\mathcal{H}$ is said to be *characteristic* with respect to $\mathcal{F}$, if it holds that $\int f dP = \int f dQ$ for any $f \in \mathcal{H}$ implies $P = Q$. Here P and Q are probability measures on $(\Omega,\mathcal{F})$.

Theorem 4. *Under Assumption 3, there exists an order between two self-adjoint conditional covariance operators $\Sigma_{YY|\boldsymbol{X}}$ and $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$ as follows.*

$$\Sigma_{YY|\boldsymbol{X}} \le \Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}. \tag{11}$$

Furthermore, if $\mathcal{H}_{\mathcal{Y}}$ is assumed to be characteristic with respect to $\mathcal{F}_{\mathcal{Y}}$, then

$$\Sigma_{YY|\boldsymbol{X}} = \Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}} \quad \text{if and only if} \quad Y \perp\!\!\!\perp \boldsymbol{X} | \sigma_{B,\boldsymbol{\phi}}. \tag{12}$$

Theorem 4 provides the theoretical underpinning for the GKDR method which will be presented in the following section. Notably, the inequality (11) in Theorem 4 indicates that the conditional covariance operator $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$ is lower bounded. Therefore, the minimization problem of $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$ with respect to the normalized generating pair $(B,\boldsymbol{\phi})$ is well-defined. Furthermore, (12) indicates that $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$ is minimized only when the associated σ-field of $B^T\boldsymbol{\phi}(X)$ is an FSDR σ-field. Therefore, we can obtain FSDR σ-fields with dimension d through the minimization of $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$ for $(B,\boldsymbol{\phi}) \in \mathbb{S}_d^p(\mathbb{R}) \times \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$. In general, Theorem 4 holds for any FSDR σ-field with dimension d no smaller than d_0. If d is greater

than d_0, Theorem 4 implies that minimizing $\Sigma^{B,\phi}_{YY|\boldsymbol{X}}$ over $\mathbb{S}^p_d(\mathbb{R}) \times \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$ leads to the generating pairs of some larger FSDR σ-fields with dimension d. If $d = d_0$, Theorem 4 implies that minimizing $\Sigma^{B,\phi}_{YY|\boldsymbol{X}}$ over $\mathbb{S}^p_{d_0}(\mathbb{R}) \times \tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})$ leads to Θ_1, the generating pairs of FSDR σ-fields with the same dimension as σ_{B^*,ϕ^*}. Since the conditional covariance operator is positive self-adjoint, minimization of the conditional covariance operator can be done by minimizing its trace. Therefore, Θ_1 can be characterized as the solution set of the following minimization problem.

$$\Theta_1 = \underset{(B,\phi):B\in\mathbb{S}^p_{d_0}(\mathbb{R}),\phi\in\tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})}{argmin} Tr[\Sigma^{B,\phi}_{YY|\boldsymbol{X}}]. \tag{13}$$

Note that Θ_1 contains infinite generating pairs, so the minimization problem (13) also have infinitely many solutions. This property will be considered in the next section when we construct the estimator for Θ_1.

3.2. Stage I of GKDR: Sample level

In this subsection, we obtain the estimator of Θ_1 using Theorem 4 and minimization problem (13), which is further referred to as the GKDR Stage I estimator. Suppose $\{(\boldsymbol{X}^{(t)}, Y^{(t)})\}_{t=1}^n$ is an independent and identically distributed (i.i.d.) sample drawn from the joint distribution $P_{\boldsymbol{X}Y}$. Let $\hat{P}_{\boldsymbol{X}Y} = \frac{1}{n}\sum_{t=1}^n \delta_{\boldsymbol{X}^{(t)}}\delta_{Y^{(t)}}$ be the empirical distribution, where $\delta_a(\cdot)$ is the Dirac delta function with point mass at a. We define the empirical cross-covariance operator $\hat{\Sigma}^{B,\phi(n)}_{Y\boldsymbol{X}} : \mathcal{H}^{B,\phi}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ as follows.

$$\langle g, \hat{\Sigma}^{B,\phi(n)}_{Y\boldsymbol{X}} f\rangle_{\mathcal{H}_{\mathcal{Y}}} = \frac{1}{n}\sum_{t=1}^n g(Y^{(t)})f(\boldsymbol{X}^{(t)}) - \big[\frac{1}{n}\sum_{t=1}^n g(Y^{(t)})\big]\big[\frac{1}{n}\sum_{t=1}^n f(\boldsymbol{X}^{(t)})\big]. \tag{14}$$

Note that the right hand side of (14) is exactly the right hand side term of (7) evaluated at the empirical distribution $\hat{P}_{\boldsymbol{X}Y}$. The empirical variance operator $\hat{\Sigma}^{B,\phi(n)}_{\boldsymbol{X}\boldsymbol{X}} : \mathcal{H}^{B,\phi}_{\mathcal{X}} \to \mathcal{H}^{B,\phi}_{\mathcal{X}}$ and $\hat{\Sigma}^{(n)}_{YY} : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ can be similarly defined.

The empirical variance operators and empirical cross-covariance operator have matrix representations. Let $\zeta^{B,\phi}_t \in \mathcal{H}^{B,\phi}_{\mathcal{X}}$ and $\eta_t \in \mathcal{H}_{\mathcal{Y}}$ $(1 \le t \le n)$ be the functions defined as follows.

$$\zeta^{B,\phi}_t = k^{B,\phi}_{\mathcal{X}}(\boldsymbol{X}^{(t)}, \cdot) - \frac{1}{n}\sum_{s=1}^n k^{B,\phi}_{\mathcal{X}}(\boldsymbol{X}^{(s)}, \cdot); \tag{15}$$

$$\eta_t = k_{\mathcal{Y}}(Y^{(t)}, \cdot) - \frac{1}{n}\sum_{s=1}^n k_{\mathcal{Y}}(Y^{(s)}, \cdot). \tag{16}$$

Let $K^{B,\phi}_{\boldsymbol{X}}$ be the Gram matrix with respect to the kernel function $k^{B,\phi}_{\mathcal{X}}$ which is defined as $(K^{B,\phi}_{\boldsymbol{X}})_{ts} = k^{B,\phi}_{\mathcal{X}}(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(s)})$, and $G^{B,\phi}_{\boldsymbol{X}}$ the centered Gram matrix defined as $G^{B,\phi}_{\boldsymbol{X}} = (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)K^{B,\phi}_{\boldsymbol{X}}(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)$, where $\mathbf{1}$ is the n-dimensional vector with all of its entries equal to 1. The Gram matrix K_Y and the centered Gram matrix G_Y with respect to the kernel function $k_{\mathcal{Y}}$ can be similarly defined. It can be verified that the matrix representation of $\hat{\Sigma}^{B,\phi(n)}_{Y\boldsymbol{X}}$ with respect

to $\{\zeta_t^{B,\boldsymbol{\phi}}\}_{t=1}^n$ and $\{\eta_t\}_{t=1}^n$ is $n^{-1}G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}}$. Similarly, the matrix representation of $\hat{\Sigma}_{\boldsymbol{X}Y}^{B,\boldsymbol{\phi}(n)}$ with respect to $\{\eta_t\}_{t=1}^n$ and $\{\zeta_t^{B,\boldsymbol{\phi}}\}_{t=1}^n$ is $n^{-1}G_Y$, the matrix representation of $\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)}$ with respect to $\{\zeta_t^{B,\boldsymbol{\phi}}\}_{t=1}^n$ is $n^{-1}G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}}$, and the matrix representation of $\hat{\Sigma}_{YY}^{(n)}$ with respect to $\{\eta_t\}_{t=1}^n$ is $n^{-1}G_Y$.

Following Fukumizu et al. [10], we define the empirical conditional covariance operator $\hat{\Sigma}_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)}$, which is the empirical version of $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$, as follows.

$$\hat{\Sigma}_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)} = \hat{\Sigma}_{YY}^{(n)} - \hat{\Sigma}_{Y\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)}(\hat{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)} + \epsilon_n I_n)^{-1}\hat{\Sigma}_{\boldsymbol{X}Y}^{B,\boldsymbol{\phi}(n)}, \tag{17}$$

where $\epsilon_n I_n$ is the regularization term required for inverting the operator. Recall that Θ_1 is characterized by the solution set of the minimization problem (13). We perform optimization at sample level by replacing the conditional covariance operator in (13) with its empirical version as follows.

$$\min_{(B,\boldsymbol{\phi}):B\in\mathbb{S}_{d_0}^p(\mathbb{R}),\boldsymbol{\phi}\in\tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})} Tr[\hat{\Sigma}_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)}] \tag{18}$$

It can be verified that the matrix representation of $\hat{\Sigma}_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)}$ with respect to $\{\eta_t\}_{t=1}^n$ is $\frac{1}{n}\big[G_Y - G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}}(G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}} + n\epsilon_n I_n)^{-1}G_Y\big]$ and its trace is

$$Tr[\hat{\Sigma}_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)}] = \frac{1}{n}Tr\big[G_Y - G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}}(G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}} + n\epsilon_n I_n)^{-1}G_Y\big] \tag{19}$$

$$= \epsilon_n Tr\big[G_Y(G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}} + n\epsilon_n I_n)^{-1}\big]. \tag{20}$$

Therefore, by replacing $\hat{\Sigma}_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}(n)}$ with its matrix representation, (18) reduces to

$$\min_{(B,\boldsymbol{\phi}):B\in\mathbb{S}_{d_0}^p(\mathbb{R}),\boldsymbol{\phi}\in\tilde{\boldsymbol{L}}^2(P_{\boldsymbol{X}})} Tr\big[G_Y(G_{\boldsymbol{X}}^{B,\boldsymbol{\phi}} + n\epsilon_n I_n)^{-1}\big]. \tag{21}$$

Notice that the first ϵ_n is omitted since the multiplier constant does not affect in the minimization problem. We denote the objective function in (21) as $GKDR(B,\boldsymbol{\phi})$.

In the objective function $GKDR(B,\boldsymbol{\phi})$, $\boldsymbol{\phi}$ is a vector of univariate transformations $\phi_1,\ldots,\phi_p$, which need to be further parametrized. Here we adopt B-splines to approximate the univariate transformations. Without loss of generality, we assume that $X_i \in [0,1]$ for $i = 1,\ldots,p$. We consider natural cubic splines of order 4. For $1 \le i \le p$ and the corresponding knots sequence $0 = t_{i,0} < t_{i,1} < \cdots < t_{i,q} = 1$, we define $S_{i,q} = \big\{\phi_i \in C^2[0,1] : \phi_i(x_i)$ is a polynomial of order 4 on each sub-interval $[t_{i,k}, t_{i,k+1}]$, and $\phi_i^{(2)}(0) = \phi_i^{(2)}(1) = \phi_i^{(3)}(0) = \phi_i^{(3)}(1) = 0\big\}$, where $C^2[0,1]$ denote twice continuously differentiable functions on $[0,1]$. Knots selection is an important issue whenever spline functions are used. In this article, we do not investigate it for our proposed method, instead we assume the knots are pre-specified using some conventional methods. For practice, we use knots placed at the sample quantiles, while for theoretical development we use equally spaced knots. $S_{i,q}$ is a q-dimensional spline

space, and we use it to approximate the univariate transformation ϕ_i. Suppose $\{B_{i,j}(x)\}_{j=1}^{q}$ is a basis of $S_{i,q}$. Denote $\boldsymbol{B}_i(x_i) = \big(B_{i,1}(x_i), \ldots, B_{i,q}(x_i)\big)$ and $\boldsymbol{B}_i(\boldsymbol{X}_i) = \big(\boldsymbol{B}_i(X_i^{(1)})^T, \ldots, \boldsymbol{B}_i(X_i^{(n)})^T\big)^T$. For convenience, the basis is chosen such that $\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{B}_i(X_i^{(t)}) = \mathbf{0}$ and $\frac{1}{n}\boldsymbol{B}_i(\boldsymbol{X}_i)^T\boldsymbol{B}_i(\boldsymbol{X}_i) = I_q$ for $i = 1, \ldots, p$. Then for any $\phi_i \in S_{i,q}$, there exists a unique column vector $\alpha \in \mathbb{R}^q$ such that $\phi_i = \boldsymbol{B}_i\alpha$. By restricting $\phi_i \in S_{i,q}$ and normalizing ϕ_i to have mean 0 and variance 1 for $i = 1, \ldots, p$, the minimization problem (21) is parametrized as follows.

$$\min_{B \in \mathbb{S}_{d_0}^{p}(\mathbb{R}), \alpha_1 \in \mathbb{R}^q, \ldots, \alpha_p \in \mathbb{R}^q} GKDR\big(B, (\boldsymbol{B}_1\alpha_1, \ldots, \boldsymbol{B}_p\alpha_p)^T\big), \tag{22}$$
$$\text{subject to } \alpha_i^T\alpha_i = 1, \quad i = 1 \ldots, p.$$

Let $(\tilde{B}_n, \hat{\alpha}_{n,1}, \ldots, \hat{\alpha}_{n,p})$ denote the solution to (22), and let $\hat{\phi}_{n,i} = \boldsymbol{B}_i\hat{\alpha}_{n,i}$ and $\hat{\boldsymbol{\phi}}_n = (\hat{\phi}_{n,1}, \ldots, \hat{\phi}_{n,p})^T$. For better visualization of the sufficient predictor $\tilde{B}_n^T\hat{\boldsymbol{\phi}}_n(\boldsymbol{X})$, we impose an orthogonal transformation P onto $\tilde{B}_n$, such that the new sufficient predictor $(\tilde{B}_nP)^T\hat{\boldsymbol{\phi}}_n(\boldsymbol{X})$ has a diagonal sample covariance matrix. This does not change the associated FSDR σ-field, since $span(\tilde{B}_n) = span(\tilde{B}_nP)$. Let $\hat{B}_n = \tilde{B}_nP$, and we refer to $(\hat{B}_n, \hat{\boldsymbol{\phi}}_n)$ as the GKDR Stage I estimator.

Recall that Θ_1 contains infinitely many generating pairs which can be characterized as the solutions to the minimization problem (13). It is not possible to estimate all the generating pairs in Θ_1 based on a finite sample, and instead only a finite number of them can be estimated. Further recall that Θ_0, which is a subset of Θ_1, is our ultimate inference target. Therefore, we need to estimate a sufficient number of generating pairs in Θ_1 at this stage, with the hope that those estimated generating pairs contain at least one pair, which can become the estimator of a generating pair in Θ_0. There is another issue we need to address when solving (22). Note that the objective function in (22) is generally nonconvex. Therefore, there may exist many local minima that are spurious in that they do not lead to good estimators of some generating pairs.

In order to estimate a sufficient number of generating pairs in Θ_1 and avoid spurious local minima, we consider the idea of multiple initializations and propose the following estimation scheme. First we randomly choose N different initializations for the minimization problem (22) to get N solutions and their associated generating pairs. Then we select the N_1 generating pairs that attain the N_1 smallest values of the GKDR objective function. The N_1 selected generating pairs are retained as the GKDR Stage I estimators, whereas the others are considered spurious solutions and discarded. In our simulation experiments and real life data applications in Section 4, we set N and N_1 to be 20 and 10, respectively. Let $\hat{\Theta}_1$ denote the collection of the N_1 GKDR Stage I estimators. Based on $\hat{\Theta}_1$, we will use conditional entropy to select a generating pair as the final estimated generating pair of the minimal σ-field σ_{B^*,ϕ^*}. This is the task that the second stage (i.e. Stage II) of the GKDR method will take on, which will be presented in Section 3.4.

3.3. Consistency for $(\hat{B}_n, \hat{\phi}_n)$

In this subsection we prove that $(\hat{B}_n, \hat{\phi}_n)$ is consistent under some conditions. The space $\mathbb{S}^p_{d_0}(\mathbb{R})$ in which B lies is naturally equipped with the geodesics distance, and we denote it as D. The formal definition of geodesics distance can be found in [15], Chapter IV. The spaces $L^2(P_{X_i})$ and $L^2(P_{\boldsymbol{X}})$ in which ϕ_i and $\boldsymbol{\phi}$ lie respectively are equipped with the L^2-type distances defined as follows.

$$L^2_{X_i}(\phi_i, \tilde{\phi}_i) = \left[E_{X_i}|\phi_i(X_i) - \tilde{\phi}_i(X_i)|^2\right]^{1/2}, \text{ for } i = 1, \ldots, p; \tag{23}$$

$$L^2_{\boldsymbol{X}}(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}) = \left[E_{\boldsymbol{X}}\left|(\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}})^T(\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}})\right|\right]^{1/2}. \tag{24}$$

Before stating the main theorem, we need some technical assumptions.

Assumption 4. *For any bounded continuous function g on $\mathcal{Y}$, the bivariate function $(B, \boldsymbol{\phi}) \mapsto E_{\boldsymbol{X}}\left[E_{Y|B^T\boldsymbol{\phi}(\boldsymbol{X})}\left(g(Y)|B^T\boldsymbol{\phi}(\boldsymbol{X})\right)^2\right]$ is continuous with respect to B and $\boldsymbol{\phi}$ equipped with the distances D^B and $L^2_{\boldsymbol{X}}$, respectively.*

Assumption 5. *There exists a measurable function $\xi : \mathcal{X} \to \mathbb{R}$ such that $E|\xi(\boldsymbol{X})|^2 < \infty$ and the Lipschitz function $\left\|k_{d_0}\left(B^T\boldsymbol{\phi}(\boldsymbol{x}), \cdot\right) - k_{d_0}\left(\tilde{B}^T\tilde{\boldsymbol{\phi}}(\boldsymbol{x}), \cdot\right)\right\|_{\mathcal{H}_{d_0}}$ $\leq \xi(\boldsymbol{X})\left(D(B, \tilde{B}) + L^2_{\boldsymbol{X}}(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}})\right)$ holds for any $(B, \boldsymbol{\phi}) \in \mathbb{S}^p_{d_0}(\mathbb{R}) \times L^2(P_{\boldsymbol{X}})$, $(\tilde{B}, \tilde{\boldsymbol{\phi}}) \in \mathbb{S}^p_{d_0}(\mathbb{R}) \times L^2(P_{\boldsymbol{X}})$, and $x \in \mathcal{X}$.*

Assumption 6. *The GKDR objective function at the population level $Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}]$ is locally convex with respect to Θ_1, that is, for any $(\tilde{B}, \tilde{\boldsymbol{\phi}}) \in \Theta_1 = \mathbb{B}^p_{d_0} \times \Phi_1$ and $\epsilon > 0$, there exists $\delta > 0$, such that if $\left|Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}] - Tr[\Sigma^{\tilde{B},\tilde{\boldsymbol{\phi}}}_{YY|\boldsymbol{X}}]\right| < \delta$, then $D(B, \mathbb{B}^p_{d_0}) < \epsilon$ and $L^2_{\boldsymbol{X}}(\boldsymbol{\phi}, \Phi_1) < \epsilon$.*

Assumption 7. *There exists a generating pair $(B, \boldsymbol{\phi}) \in \Theta_1$ such that ϕ_i is twice continuously differentiable for $i = 1, \ldots, p$.*

Assumptions 4 and 5 are the generalizations of Assumptions (A-1) and (A-3) in [10] for proving the consistency of kernel dimension reduction. Assumption 6 is similar to Assumption A1 in [22]. Assumption 7 is to ensure that the univariate transformations $\phi_1, \ldots, \phi_p$ can be approximated uniformly by the spline functions with certain precision.

Theorem 5. *Suppose Assumptions 3-7 hold, and the kernel function k_{d_0} is continuous and bounded. For any open set $\tilde{\mathbb{B}}_1 \supseteq \mathbb{B}^p_{d_0}$ and any open set $\tilde{\Phi}_1 \supseteq \Phi_1$, one can use spline functions with sufficiently high order q such that, if the regularization parameter ϵ_n satisfies that*

$$\epsilon_n \to 0 \text{ and } n^{1/2}\epsilon_n \to \infty \quad (n \to \infty), \tag{25}$$

then $\lim_{n\to\infty} \Pr(\hat{B}_n \in \tilde{\mathbb{B}}_1, \hat{\boldsymbol{\phi}}_n \in \tilde{\Phi}_1) = 1$.

Note that we cannot ensure that as n goes to infinity, $(\hat{B}_n, \hat{\boldsymbol{\phi}}_n)$ converge to a fixed generating pair $(B_0, \boldsymbol{\phi}_0) \in \mathbb{B}^p_{d_0} \times \Phi_1$ in probability, whereas, as n goes to infinity, we can always find a generating pair $(B_n, \boldsymbol{\phi}_n) \in \mathbb{B}^p_{d_0} \times \Phi_1$ such that $(\hat{B}_n, \hat{\boldsymbol{\phi}}_n)$ is close enough to $(B_n, \boldsymbol{\phi}_n)$ in probability.

3.4. Stage II of GKDR and conditional entropy

In this subsection, we will focus on estimating Θ_0 and obtaining the GKDR Stage II estimator. Recall that $\Theta_0 = \mathbb{B}^p_{d_0} \times \Phi_0$ and $\Theta_1 = \mathbb{B}^p_{d_0} \times \Phi_1$, and the $\phi_i^{*'}s$ in Φ_0 are coarser than the $\phi_i's$ in Φ_1 for $i \in \mathcal{I}_{B^*}$, that is, $\sigma\big(\phi_i^*(X_i)\big) \subseteq \sigma\big(\phi_i(X_i)\big)$ for $i \in \mathcal{I}_{B^*}$. Here we will show that the conditional entropy of $\phi_i^{*'}s$ in Φ_0 are no smaller than that of $\phi_i's$ in Φ_1, so we can obtain Θ_0 by maximizing the conditional entropy over Θ_1. Following Cover and Thomas [6], for a univariate random variable X and a univariate transformation $\phi(X)$, we define the *conditional entropy* of X given $\phi(X)$, denoted as $H\big[X|\phi(X)\big]$, as follows.

$$\begin{aligned} H\big[X|\phi(X)\big] &= -E\big[\log p(X|\phi(X))\big] \\ &= -E_{\phi(X)}\Big[E_{X|\phi(X)}\big[\log p(X|\phi(X))\big]\Big], \end{aligned}$$

where $p(X|\phi(X))$ denotes the conditional distribution of X given $\phi(X)$. It can be shown that the conditional entropy $H\big[X|\phi(X)\big] \geq 0$, and the equality holds if and only if $\sigma(X) = \sigma\big(\phi(X)\big)$. For a random vector $\boldsymbol{X}$ and a generating pair $(B, \boldsymbol{\phi})$, the B-weighted conditional entropy of $\boldsymbol{X}$ given $\boldsymbol{\phi}(\boldsymbol{X})$, denoted as $H_B\big[\boldsymbol{X}|\boldsymbol{\phi}(\boldsymbol{X})\big]$, is defined as follows.

$$H_B\big[\boldsymbol{X}|\boldsymbol{\phi}(\boldsymbol{X})\big] = \sum_{i=1}^{p} w_i H\big[X_i|\phi_i(X_i)\big],$$

where $w_i = \|B_{i,\cdot}\|_2$ is the L_2 norm of the i-th row of matrix B. In fact, Θ_0 can be characterized as the solution set of a maximization problem of $H_B[\boldsymbol{X}|\boldsymbol{\phi}(\boldsymbol{X})]$ over Θ_1, which will be presented in Theorem 6. Before that, we require an assumption to ensure that the conditional entropy $H_{B^*}\big[\boldsymbol{X}|\boldsymbol{\phi}^*(\boldsymbol{X})\big]$ is invariant in Θ_0, that is, for any two generating pairs $(B^*, \boldsymbol{\phi}^*)$ and $(\tilde{B}, \tilde{\boldsymbol{\phi}})$ in Θ_0, we have $H_{B*}\big[\boldsymbol{X}|\boldsymbol{\phi}^*(\boldsymbol{X})\big] = H_{\tilde{B}}\big[\boldsymbol{X}|\tilde{\boldsymbol{\phi}}(\boldsymbol{X})\big]$.

Assumption 8. *There exists a generating pair $(B^*, \boldsymbol{\phi}^*) \in \Theta_0$ such that the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{\phi}^*(\boldsymbol{X})$, denoted as $p\big(\boldsymbol{X}|\boldsymbol{\phi}^*(\boldsymbol{X})\big)$, is discrete.*

Assumption 8 is satisfied if there exists a generating pair $(B^*, \boldsymbol{\phi}^*) \in \Theta_0$ such that for any real number c and $1 \leq i \leq p$, the level set of $\boldsymbol{\phi}_i^*$, defined as $L_c(\boldsymbol{\phi}_i^*) = \{x : \boldsymbol{\phi}_i^*(x) = c\}$, is countable.

Theorem 6. *Suppose X is a univariate random variable and ψ and ϕ are two univariate transformations. If $\sigma(X) \supseteq \sigma\big(\psi(X)\big) \supseteq \sigma\big(\phi(X)\big)$ and the conditional*

distribution $p\big(X|\phi(X)\big)$ *is discrete, then the conditional entropies satisfy the following decomposition rule.*

$$H\big[X|\phi(X)\big] = H\big[X|\psi(X)\big] + H\big[\psi(X)|\phi(X)\big]. \tag{26}$$

Furthermore, suppose Assumption 8 holds and for any $(B^*, \boldsymbol{\phi}^*) \in \Theta_0$ *and* $(\tilde{B}, \tilde{\boldsymbol{\phi}}) \in \Theta_1$*, one has*

$$H_{B^*}\big[\boldsymbol{X}|\boldsymbol{\phi}^*(\boldsymbol{X})\big] \geq H_{\tilde{B}}\big[\boldsymbol{X}|\tilde{\boldsymbol{\phi}}(\boldsymbol{X})\big], \tag{27}$$

and the equivalence holds if and only if $(\tilde{B}, \tilde{\boldsymbol{\phi}}) \in \Theta_0$.

A direct conclusion of Theorem 6 is that, among all the generating pairs in Θ_1, the conditional entropy $H_B\big[\boldsymbol{X}|\boldsymbol{\phi}(\boldsymbol{X})\big]$ attains maximum at $(B^*, \boldsymbol{\phi}^*)$ in Θ_0, that is,

$$\Theta_0 = \underset{(B,\boldsymbol{\phi})\in\Theta_1}{argmax}\ H_B\big[\boldsymbol{X}|\boldsymbol{\phi}(\boldsymbol{X})\big]. \tag{28}$$

Therefore, we can obtain the generating pairs in Θ_0 by selecting those in Θ_1 such that $H_B\big[\boldsymbol{X}|\boldsymbol{\phi}(\boldsymbol{X})\big]$ is maximized. This is referred to as Stage II of GKDR at population level.

To obtain the estimator of Θ_0, we choose the generating pair in $\hat{\Theta}_1$ such that the empirical conditional entropy is maximized. The empirical conditional entropy is evaluated by slicing the data. Let $\hat{H}\big[X_i|\hat{\phi}_i(X_i)\big]$ and $\hat{H}_{\hat{B}}\big[\boldsymbol{X}|\hat{\boldsymbol{\phi}}(\boldsymbol{X})\big]$ denote the empirical version of $H\big[X_i|\phi_i(X_i)\big]$ and $H_B\big[\boldsymbol{X}|\boldsymbol{\phi}(\boldsymbol{X})\big]$, respectively. For component X_i, suppose we have n i.i.d samples $X_i^{(1)}, \ldots, X_i^{(n)}$. We divide the range of $X_i^{(t)}$ into W equal-sized slices $S_1, \ldots, S_W$, and the range of $\hat{\phi}_i(X_i^{(t)})$ into W equal-sized slices $T_1, \ldots, T_W$. Let n_{uv} be the number of data points such that $X_i^{(t)}$ falls into S_u and $\hat{\phi}_i(X_i^{(t)})$ falls into T_v, and $n_{\cdot v}$ the number of data points such that $\hat{\phi}_i(X_i^{(t)})$ falls into T_v. The empirical conditional entropies are given as

$$\begin{aligned}\hat{H}\big[X_i|\hat{\phi}_i(X_i)\big] =& \sum_{v=1}^{W} \frac{n_{\cdot j}}{n} \sum_{u=1}^{W} \frac{n_{uv}}{n_{\cdot v}} \log \frac{n_{uv}}{n_{\cdot v}} \\ =& \frac{1}{n} \sum_{v=1}^{W} \sum_{u=1}^{W} n_{uv} \log \frac{n_{uv}}{n_{\cdot v}},\end{aligned}$$

and $\hat{H}_{\hat{B}}\big[\boldsymbol{X}|\hat{\boldsymbol{\phi}}(\boldsymbol{X})\big] = \sum_{i=1}^{p} \hat{w}_i \hat{H}\big[X_i|\hat{\phi}_i(X_i)\big]$, where $\hat{w}_i = \|\hat{B}_{i,\cdot}\|_2$. Then the solution that maximizes $\hat{H}_{\hat{B}}\big[\boldsymbol{X}|\hat{\boldsymbol{\phi}}(\boldsymbol{X})\big]$ over $\hat{\Theta}_1$ is defined as the GKDR Stage II estimator, and is considered as an approximation of one representing generating pair in Θ_0.

4. Implementation and numerical results

4.1. Low rank approximation

The GKDR method also suffers from the issue of computational inefficiency as other kernel methods do. Various methods have been proposed to address this

issue in the literature, among which low rank matrix approximation achieves much success. In general, the Gram matrices in the kernel methods have fast decreasing eigen values, henceforth they can be approximated by matrices with much smaller rank of $m << n$. The Nystrom method [25] is one of the widely used low rank matrix approximation approach, and we incorporate it into the GKDR method to improve the computational efficiency. The Nystrom method works as follows. First, it chooses m columns as the *pivots* of the centered Gram matrix $G_X^{B,\boldsymbol{\phi}}$. Denote $G_{n,m} \in \mathbb{R}^{n\times m}$ as the submatrix of $G_X^{B,\boldsymbol{\phi}}$ whose columns are in the pivot set, and $G_{m,m} \in \mathbb{R}^{m\times m}$ as the submatrix of $G_X^{B,\boldsymbol{\phi}}$ whose rows and columns are both in the pivot set. Then $\tilde{G} = G_{n,m}G_{m,m}^{-1}G_{n,m}^T$ is the approximation of the original Gram matrix $G_X^{B,\boldsymbol{\phi}}$, reducing the computation complexity to $O(m^2 n)$.

Furthermore, we use the incomplete Cholesky decomposition method [9] to choose the pivot set in this article. Basically, the method operates by iteratively choosing the columns until the difference between $\tilde{G}$ and $G_X^{B,\boldsymbol{\phi}}$ is small enough, i.e. $\|G_X^{B,\boldsymbol{\phi}} - \tilde{G}\|_1 \leq tol$, where tol is a pre-specified small constant. After replacing $G_X^{B,\boldsymbol{\phi}}$ with $\tilde{G}$, the inverse term $(G_X^{B,\boldsymbol{\phi}} + n\epsilon_n I_n)^{-1}$ in the objective function in (21) can be approximated by $(G_{n,m}G_{m,m}^{-1}G_{n,m}^T + \epsilon I_n)^{-1}$. By the Sherman-Morrison-Woodbury (SMW) formula, which is

$$(D + VV^T)^{-1} = D^{-1} - D^{-1}V(I + V^T D^{-1}V)^{-1}V^T D^{-1},$$

we transform $(G_{n,m}G_{m,m}^{-1}G_{n,m}^T + \epsilon I_n)^{-1}$ to $I_n - G_{n,m}(G_{n,m}^T G_{n,m} + \epsilon_n G_{m,m})^{-1} G_{n,m}^T$. Henceforth, the objective function in (21) is reduced to

$$GKDR(B, \boldsymbol{\phi}) \approx Tr\Big[G_Y\big(I_n - G_{n,m}(G_{n,m}^T G_{n,m} + \epsilon_n G_{m,m})^{-1} G_{n,m}^T\big)\Big]. \quad (29)$$

Because the term involving inversion in (29) is an $m \times m$ matrix, its computation complexity is reduced from $O(n^3)$ to $O(m^3 + n^2)$ for one evaluation of the objective function.

4.2. Determination of dimension d_0

As mentioned previously, we assume d_0, the dimension of the minimal FSDR σ-field is known a priori, while in practice we always need to determine one operable d_0. Note that if we misspecify d_0 as a larger integer d, Theorem 4 guarantees that the generating pair $(B, \boldsymbol{\phi})$ estimated by minimizing $\Sigma_{YY|\boldsymbol{X}}^{B,\boldsymbol{\phi}}$, is still associated with an FSDR σ-field, although this σ-field is larger than the minimal FSDR σ-field. Leveraging this property, we propose a heuristic method to determine the dimension d_0. We first use the χ^2 statistic method [17] to estimate d_1, the dimension of the central subspace under linear SDR, which is no less than d_0. Then for each $d \leq d_1$, we apply the GKDR method and obtain the corresponding generating pair estimate $(\hat{B}_d, \hat{\boldsymbol{\phi}}_d)$. For $d \geq d_0$, the σ-field associated with $(\hat{B}_d, \hat{\boldsymbol{\phi}}_d)$ is an FSDR σ-field and is sufficient in predicting Y; whereas for $d < d_0$, the σ-field associated with $(\hat{B}_d, \hat{\boldsymbol{\phi}}_d)$ is not. Therefore, d_0

is the smallest d such that the σ-field associated with $(\hat{B}_d, \hat{\boldsymbol{\phi}}_d)$ is sufficient in predicting Y. For each $d \leq d_1$, we fit a linear regression model between Y and $\hat{B}_d^T \hat{\boldsymbol{\phi}}_d(\boldsymbol{X})$, and denote the R-squared value of the regression by R_d^2. If R_{d-1}^2 is approximately equal to R_d^2 in the sense that $\frac{R_d^2 - R_{d-1}^2}{R_{d-1}^2} \leq 0.05$, we consider the capabilities of $\hat{B}_d^T \hat{\boldsymbol{\phi}}_d(\boldsymbol{X})$ and $\hat{B}_{d-1}^T \hat{\boldsymbol{\phi}}_{d-1}(\boldsymbol{X})$ in predicting Y to be almost the same, which indicates that $d-1$ is more preferable than d. By decreasing d from d_1 one at a time, we find the largest d such that R_d^2 is not approximately equal to R_{d-1}^2, and use this d as a reasonable estimator of d_0.

4.3. Numerical implementation scheme

A scheme for deriving a representative generating pair of the minimal FSDR σ-field is shown below.

1 For $i = 1, \ldots, p$, take the sample quantiles as knots and obtain the normalized natural cubic spline basis $\boldsymbol{B}_{i,q}(x_i) = \big(B_{i,1}(x_i), \ldots, B_{i,q}(x_i)\big)^T$ for estimating $\phi_i(x_i)$.

2 Solve the following minimization problem via gradient descent method on matrix manifolds [24]:

$$\min_{B \in \mathbb{S}_{d_0}^p(\mathbb{R}), \alpha_1 \in \mathbb{R}^q, \ldots, \alpha_p \in \mathbb{R}^q} Tr\Big[G_Y\big(I_n - G_{n,m}(G_{n,m}^T G_{n,m} + \epsilon_n G_{m,m})^{-1} G_{n,m}^T\big)\Big]$$
$$\text{subject to } \alpha_i^T \alpha_i = 1, \quad i = 1 \ldots, p,$$

where G is the centered kernel matrix $G_X^{B,\boldsymbol{\phi}}$ with $\phi_i(x_i) = \boldsymbol{B}_{i,q}(x_i)\alpha_i$, and $G_{n,m}$ is the submatrix of G whose columns are in the pivot set, $G_{m,m}$ is the submatrix of G whose rows and columns are in the pivot set. The pivot set is determined by iteratively choosing the columns until $\|G - G_{n,m}G_{m,m}^{-1}G_{n,m}^T\|_1 \leq tol$. ϵ_n is fixed as 0.001 and tol is fixed as a relatively small number, i.e. 10^{-8}.

3 Repeat the last step for N times with random initialization and obtain N generating pairs $\big\{(\hat{B}_{(l)}, \hat{\boldsymbol{\phi}}_{(l)})\big\}_{l=1}^N$. Choose top N_1 with the smallest GKDR objective function value in these N generating pairs, and let $\hat{\Theta}_1$ denote their collection. We determine $N = 20$ and $N_1 = 10$ in practice.

4 Let $(\hat{B}_0, \hat{\boldsymbol{\phi}})$ be the one that maximizes $\hat{H}\big[\boldsymbol{X} | \boldsymbol{\phi}(\boldsymbol{X})\big]$ in $\hat{\Theta}_1$.

5 Perform eigenvalue decomposition: $\hat{B}_0^T \hat{\boldsymbol{\phi}}(X) \hat{\boldsymbol{\phi}}(X)^T \hat{B}_0 = PDP^T$, where P is orthogonal matrix and D is diagonal matrix. Let $\hat{B} = \hat{B}_0 P$, and the output $(\hat{B}, \hat{\boldsymbol{\phi}})$ is called the GKDR Stage II estimator.

4.4. Simulation results

In this subsection we evaluate the performance of our proposed GKDR method via simulation studies. We use the Gaussian radial basis function kernel $k(x, y) = \exp(-\frac{\|x-y\|_2^2}{c})$ with $c = 10$, and we fix the regularization parameter ϵ_n to be 10^{-4}

TABLE 1
Pearson correlation coefficients of the first three components in Models 1-4.

	$cor(\phi_1(X_1), \hat{\phi}_1(X_1))$				$cor(\phi_2(X_2), \hat{\phi}_2(X_2))$				$cor(\phi_3(X_3), \hat{\phi}_3(X_3))$			
	GKDR		FDR		GKDR		FDR		GKDR		FDR	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Model 1	0.901	0.022	0.823	0.101	0.904	0.114	0.873	0.084	0.745	0.068	0.623	0.132
Model 2	0.949	0.022	0.926	0.014	0.910	0.024	0.943	0.021	—	—	—	—
Model 3	0.898	0.022	0.588	0.277	0.894	0.037	0.893	0.027	—	—	—	—
Model 4	0.901	0.033	0.745	0.047	—	—	—	—	—	—	—	—

TABLE 2
Pearson correlation coefficients of the first ten components in Model 5.

	GKDR		FDR	
	mean	sd	mean	sd
$cor(\phi_1(X_1), \hat{\phi}_1(X_1))$	0.896	0.088	0.536	0.343
$cor(\phi_2(X_2), \hat{\phi}_2(X_2))$	0.964	0.036	0.808	0.123
$cor(\phi_3(X_3), \hat{\phi}_3(X_3))$	0.895	0.088	0.963	0.047
$cor(\phi_4(X_4), \hat{\phi}_4(X_4))$	0.892	0.067	0.971	0.066
$cor(\phi_5(X_5), \hat{\phi}_5(X_5))$	0.885	0.052	0.950	0.042
$cor(\phi_6(X_6), \hat{\phi}_6(X_6))$	0.877	0.107	0.964	0.089
$cor(\phi_7(X_7), \hat{\phi}_7(X_7))$	0.923	0.057	0.607	0.325
$cor(\phi_8(X_8), \hat{\phi}_8(X_8))$	0.911	0.083	0.620	0.302
$cor(\phi_9(X_9), \hat{\phi}_9(X_9))$	0.703	0.200	0.834	0.170
$cor(\phi_{10}(X_{10}), \hat{\phi}_{10}(X_{10}))$	0.780	0.153	0.532	0.306

and the tolerance for low rank approximation *tol* to be 10^{-8}. To estimate the transformations, we choose the natural cubic splines with knots placed at the corresponding sample quantiles. Six knots are used in the simulation studies. We consider five different models in simulation studies. Model 1 is the example discussed in Section 1, Model 2 has lower dimensionality under the FSDR framework than the original SDR framework, Model 3 has multiple FSDR σ-fields achieving the smallest dimensionality, and Model 4 considers a non-regression setting. In Models 1-4, the dimensions of $\boldsymbol{X}$ are 10, while in Model 5, $\boldsymbol{X}$ is 30-dimensional.

We perform simulations to evaluate the performance of dimension reduction. First, we use the Pearson correlation coefficient between the estimated transformations and the true transformations to measure the partial performance of dimension reduction. We compare the Pearson correlation coefficients of our proposed GKDR method with FDR [23]. The means and standard deviations of the Pearson correlation coefficients are shown in Tables 1-2. Further, we use the RV coefficient [20] between the estimated sufficient predictors and the true sufficient predictors under flexible SDR to measure the overall performance of dimension reduction. RV coefficient is a multivariate generalization of squared Pearson correlation coefficient and is able to measure the similarity between two random vectors. We compare the RV coefficients of the GKDR method with FDR and three other popular dimension reduction methods which are KDR [10], MAVE [27] and KSIR [26]. The means and standard deviations of the RV coefficients are shown in Table 3. For each model, we repeat the simulation study for 100

TABLE 3
RV coefficients for five methods.

	GKDR		FDR		KDR		MAVE		KSIR	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Model 1	0.451	0.079	0.380	0.069	0.025	0.013	0.020	0.009	0.012	0.013
Model 2	0.849	0.029	0.871	0.027	0.105	0.105	0.042	0.078	0.126	0.079
Model 3	0.763	0.039	0.589	0.138	0.149	0.052	0.169	0.034	0.139	0.042
Model 4	0.655	0.090	0.546	0.068	0.268	0.155	0.335	0.064	0.301	0.064
Model 5	0.571	0.155	0.295	0.070	0.048	0.040	0.295	0.075	0.105	0.056

times with 500 data points in each replicate.

- Model 1

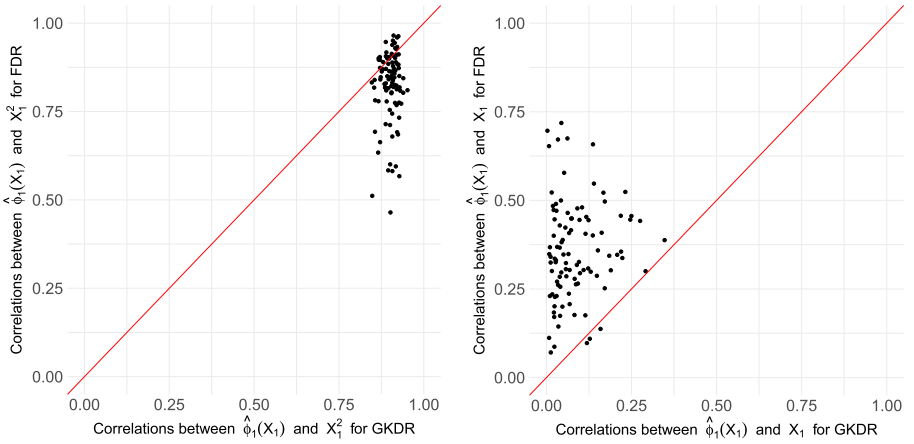
$$Y = X_1^2/(\sin(X_2) + X_3^2 + 2) + \epsilon, \tag{30}$$

where $p = 10, \boldsymbol{X} \sim \frac{1}{2}N(-\mathbf{1}_{10}, \frac{1}{4}I_{10}) + \frac{1}{2}N(\mathbf{1}_{10}, \frac{1}{4}I_{10})$ is the ten-dimensional Gaussian mixture explanatory vector, and ϵ follows $N(0, 0.25)$ and is independent with $\boldsymbol{X}$. Under the FSDR framework, we can construct a representative generating pair of the minimal FSDR σ-field as follows. $\phi_1^*(X_1) = X_1^2$, $\phi_2^*(X_2) = \sin(X_2)$, $\phi_3^*(X_3) = X_3^2$, $\phi_i^*(X_i) = 0$ for $i = 4, \ldots, 10$, and $B^* = (e_1, e_2 + e_3)$. The sufficient predictor is $(X_1^2, \sin(X_2) + X_3^2)^T$, and the dimension of the minimal FSDR σ-field is 2. We also set the dimensionality to be 2 for the other methods. Table 1 demonstrates that GKDR accurately recovers the univariate transformations and achieves much better performance in the estimation for the first and the third components. This echoes the statement mentioned in Section 1 that the dimension of B is insufficient to characterize flexible SDR. In this example, we can construct another generating pair as follows. $\tilde{\phi}_i(X_1) = X_1$, $\tilde{\phi}_2(X_2) = \sin(X_2)$, $\tilde{\phi}_3(X_3) = X_3^2$, $\tilde{\phi}_i(X_i) = 0$ for $i = 4, \ldots, 10$, and $\tilde{B} = (e_1, e_2 + e_3)$. The only difference between $(B^*, \boldsymbol{\phi}^*)$ and $(\tilde{B}, \tilde{\boldsymbol{\phi}})$ is the transformation of the first component. The dimension of $\tilde{B}$ is still equal to 2, but $\sigma_{\tilde{B},\tilde{\boldsymbol{\phi}}}$ is larger than $\sigma_{B^*,\boldsymbol{\phi}^*}$ and FDR cannot discriminate between $B^{*T}\boldsymbol{\phi}^*(\boldsymbol{X})$ and $\tilde{B}^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$. Indeed, from Figure 1 we see that $\hat{\phi}_1$ estimated by GKDR demonstrates higher correlation with X_1^2, whereas $\hat{\phi}_1$ estimated by FDR demonstrates higher correlation with X_1. Furthermore, we compare the RV coefficients of the five methods. Table 3 shows that both GKDR and FDR work well, whereas KDR, MAVE and KSIR fail to identify the true sufficient predictor. This is expected, since KDR, MAVE and KSIR are methods under linear SDR model and do not estimate the univariate transformations.

- Model 2

$$Y = \sin(3X_1) + 0.5 * (X_2 - 0.5)^2 + \epsilon, \tag{31}$$

where $p = 10, \boldsymbol{X} \sim N(0, I_{10})$ is the ten-dimensional independent Gaussian explanatory vector, and ϵ follows $N(0, 0.25)$ and is independent with $\boldsymbol{X}$. Under the FSDR framework, we can construct a representative generating pair of the minimal FSDR σ-field as follows. $\phi_1^*(X_1) = \sin(3X_1)$, $\phi_2^*(X_2) = 0.5*(X_2-0.5)^2$, $\phi_i^*(X_i) = 0$ for $i = 3, \ldots, 10$, and $B^* = e_1 + e_2$. In this example, the sufficient predictor is $\sin(3X_1)+0.5*(X_2-0.5)^2$, and the dimension of the minimal FSDR



(a) (b)

FIG 1. *Correlations between $\hat{\phi}_1(X_1)$ and X_1^2 and between $\hat{\phi}_1(X_1)$ and X_1 for GKDR and FDR. Each point represents a simulation replicate.*

σ-field is 1. We set the dimensionality to be 1 for the other methods. From Tables 1 and 3 we see that GKDR and FDR accurately recover the univariate transformations and outperform the other three methods. Note that in this example there does not exist other FSDR σ-fields with the same dimensions as the minimal FSDR σ-field, so FDR achieves slightly larger RV coefficient than GKDR.

- Model 3

$$Y = \sin(3X_1) * (X_2 + 0.5)^2 + \epsilon, \tag{32}$$

where $p = 10, \boldsymbol{X} \sim N(0, 0.6 * I_{10} + 0.4 * \mathbf{1}_{10}\mathbf{1}_{10}^T)$ is the ten-dimensional Gaussian explanatory vector, and ϵ follows $N(0, 0.25)$ and is independent with $\boldsymbol{X}$. A representative generating pair of the minimal FSDR σ-field can be constructed as follows. $\phi_1^*(X_1) = \sin(3X_1)$, $\phi_2^*(X_2) = (X_2 + 0.5)^2$, $\phi_i^*(X_i) = 0$ for $i = 3, \ldots, 10$, and $B^* = (e_1, e_2)$. In this example, the sufficient predictor is $\big(\sin(3X_1), (X_2 + 0.5)^2\big)^T$, and the dimension of the minimal FSDR σ-field is 2. We set the dimensionality to be 2 for the other methods. Again from Tables 1 and 3, we see that GKDR and FDR accurately recover the univariate transformations and outperform the other three methods. In particular, GKDR achieves higher correlation and smaller standard deviation when estimating the first univariate transformation compared to FDR in this example. The reason is the same as shown in Model 1 that the dimension of B is insufficient to characterize flexible SDR. In this specific example, another generating pair of an FSDR σ-field with the same dimension as $\sigma_{B^*,\boldsymbol{\phi}^*}$ can be constructed as follows. $\tilde{\phi}_i(X_i) = X_i$ for $i = 1, 2$, $\tilde{\phi}_i(X_i) = 0$ for $i = 3, \ldots, 10$, and $\tilde{B} = (e_1, e_2)$. The dimensions of B^* and $\tilde{B}$ are both equal to 2, but $\sigma_{\tilde{B},\tilde{\boldsymbol{\phi}}}$ is larger than $\sigma_{B^*,\boldsymbol{\phi}^*}$ and FDR cannot discriminate between $B^{*T}\boldsymbol{\phi}^*(\boldsymbol{X})$ and $\tilde{B}^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$.

- Model 4

$$Y = (X_1 - 0.5)^2 * \epsilon, \tag{33}$$

where $p = 10, \boldsymbol{X} \sim N(0, I_{10})$ is the ten-dimensional Gaussian explanatory vector, and ϵ follows $N(0, 0.25)$ and is independent with $\boldsymbol{X}$. A representative generating pair of the minimal FSDR σ-field can be constructed as follows. $\phi_1^*(X_1) = (X_1 - 0.5)^2$, $\phi_i^*(X_i) = 0$ for $i = 2, \ldots, 10$, and $B^* = e_1$. In this example, the sufficient predictor is $(X_1 - 0.5)^2$, and the dimension of the minimal FSDR σ-field is 1. We set the dimensionality to be 1 for the other methods. From Tables 1 and 3, we see that GKDR accurately recovers the univariate transformation and outperforms the other methods. In particular, Table 1 demonstrates that GKDR achieves better performance in the estimation of univariate transformations than FDR. In this specific example, another generating pair of an FSDR σ-field with the same dimension as $\sigma_{B^*, \boldsymbol{\phi}^*}$ can be constructed as follows. $\tilde{\phi}_1(X_1) = X_1$, $\tilde{\phi}_i(X_i) = 0$ for $i = 2, \ldots, 10$, and $\tilde{B} = e_1$. The dimensions of B^* and $\tilde{B}$ are both equal to 1, but $\sigma_{\tilde{B}, \tilde{\boldsymbol{\phi}}}$ is larger than $\sigma_{B^*, \boldsymbol{\phi}^*}$ and FDR cannot discriminate between $B^{*T}\boldsymbol{\phi}^*(\boldsymbol{X})$ and $\tilde{B}^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})$.

- Model 5

$$\begin{aligned} Y = &\Big[-2 + \frac{2}{1 + 25X_1^2} + \sin(\pi X_2) + \exp(X_3) + X_4 + (X_5 - 0.5)^2 \Big] * \\ &\Big[-3 + \frac{4}{1 + \exp(X_6)} + \cos(\pi X_7) + \log(1 + 9X_8^2) + X_9^3 + 2X_{10}^4 \Big] + \epsilon, \end{aligned} \tag{34}$$

where $p = 30, \boldsymbol{X} \sim Unif[-1, 1]^{30}$, and ϵ follows $N(0, 0.01)$ and is independent with $\boldsymbol{X}$. A representative generating pair of the minimal FSDR σ-field can be constructed as follows. $\phi_1^*(X_1) = 2/(1 + 25X_1^2)$, $\phi_2^*(X_2) = \sin(\pi X_2)$, $\phi_3^*(X_3) = \exp(X_3)$, $\phi_4^*(X_4) = X_4$, $\phi_5^*(X_5) = (X_5 - 0.5)^2$, $\phi_6^*(X_6) = 4/(1 + \exp(X_6))$, $\phi_7^*(X_7) = \cos(\pi X_7)$, $\phi_8^*(X_8) = \log(1 + 9X_8^2)$, $\phi_9^*(X_9) = X_9^3$, $\phi_{10}^*(X_{10}) = 2X_{10}^4$, $\phi_i^*(X_i) = 0$ for $i = 11, \ldots, 30$, $B^* = (\sum_{i=1}^5 e_i, \sum_{i=6}^{10} e_i)$. In this example, the sufficient predictor is $\big(2/(1+25X_1^2) + \sin(\pi X_2) + \exp(X_3) + X_4 + (X_5 - 0.5)^2, 4/(1 + \exp(X_6)) + \cos(\pi X_7) + \log(1 + 9X_8^2) + X_9^3 + 2X_{10}^4\big)^T$, and the dimension of the minimal FSDR σ-field is 2. We set the dimensionality to be 2 for the other methods. From Table 2 we see that the correlation coefficients of all ten univariate transformations are larger than 0.7 for GKDR, while the correlation coefficients of the first, seventh, eighth and tenth univariate transformations are smaller than 0.7 for FDR. Table 3 shows that GKDR works much better than FDR under this sophisticated model, and the linear SDR methods including KDR, MAVE and KSIR fail to identify the true sufficient predictors.

In the simulation study above, we have assumed that the true dimension d_0 of each model's minimal FSDR σ-fields is known. Recall that in Section 4.2, we proposed a heuristic method for determining d_0 in practice. We further conducted a simulation study of the effectiveness of this heuristic method under Models 1 – 5. For each model, we generated 100 samples, and then applied the heuristic method to each sample with d_1 set to be four. The proportions

TABLE 4
Determination of the true dimensions (i.e., d_0's) of Models 1-5 using the heuristic method. For each model, the proportions of the times that different dimensions were selected by the heuristic method are shown, and the highest proportion is highlighted in bold.

	d_0	selected dimension			
		1	2	3	4
Model 1	2	27%	**35%**	15%	23%
Model 2	1	**98%**	0%	0%	2%
Model 3	2	15%	**70%**	6%	9%
Model 4	1	4%	**73%**	12%	11%
Model 5	2	11%	**37%**	28%	24%

of the times that the heuristic method selected the candidate dimensions are presented in Table 4. From the table, the heuristic method selected the true dimensions of Models 2 and 3 with proportions 98% and 70%, respectively. For Models 4 and 5, the heuristic method tends to over-estimate the true dimension. The total proportions of correctly estimating d_0 plus over estimating d_0 by only one are 77% and 65% for Models 4 and 5, respectively. We argue that slight over-estimation in those two models is still acceptable, because the dimensions have been substantially reduced, and the sufficient directions have been retained. The heuristic method under-estimated the true dimension of Model 1 with proportion 27%. This may not be acceptable in practice, because under-estimation leads to loss of information. This model clearly presented challenges for the heuristic method. The determination of the true dimension of the minimal FSDR σ-field is an important and challenging problem. More research on this problem is needed in the future.

4.5. PM2.5 Data

We apply the GKDR method to the PM2.5 dataset originally analyzed in Liang et al. [19]. This dataset is used to study the relationship between the PM2.5 concentration level and the meteorological conditions in five cities in China from January 1, 2010 to December 31, 2015, and it contains 52584 hourly measured instances during that period of time. In our study, we choose the dataset of PM2.5 in Beijing, where the air pollution is most severe, to perform statistical analysis. In this dataset, each instance consists of five time-related attributes including $YEAR$, $MONTH$, DAY, $SEASON$, and $HOUR$, PM2.5 readings measured at the US Embassy in Beijing, and eight other meteorological attributes, which are dew point ($DEWP$), relative humidity ($HUMI$), air pressure ($PRES$), temperature ($TEMP$), combined wind direction ($CBWD$), cumulated wind speed (IWS), hourly precipitation ($PREC$), and cumulated precipitation ($IPREC$). Here $IPREC$ is redundant in that it can be derived from $PREC$, $CBWD$ is categorical with only five levels, and $PREC$ is highly imbalanced with over 95% zero values. Therefore, we remove these three attributes in the subsequent analysis. The PM2.5 reading is considered to be the response, and the remaining five meteorological attributes are considered to be the explanatory variables.

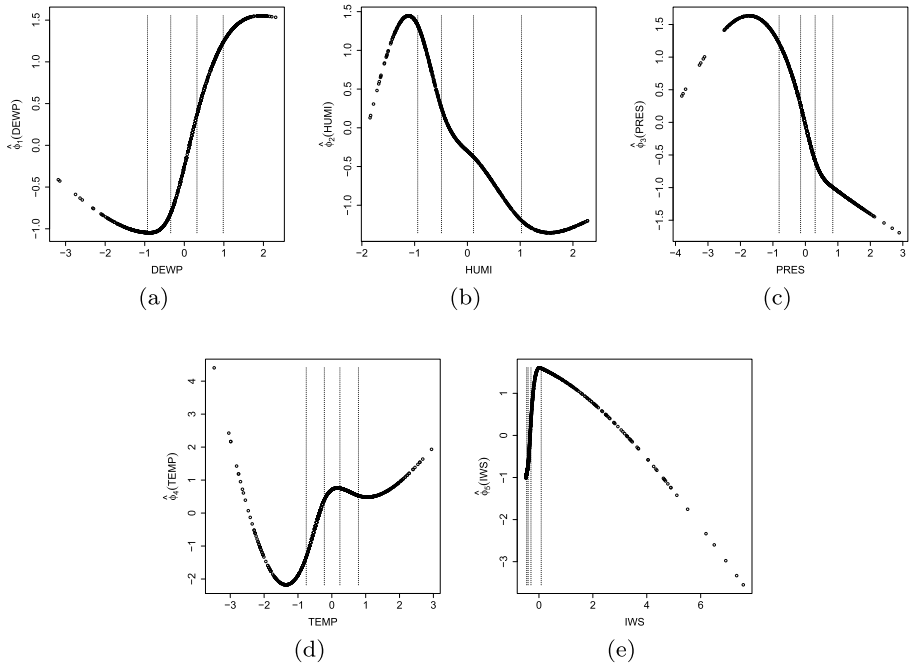


FIG 2. *Estimated univariate transformations for PM 2.5 data.*

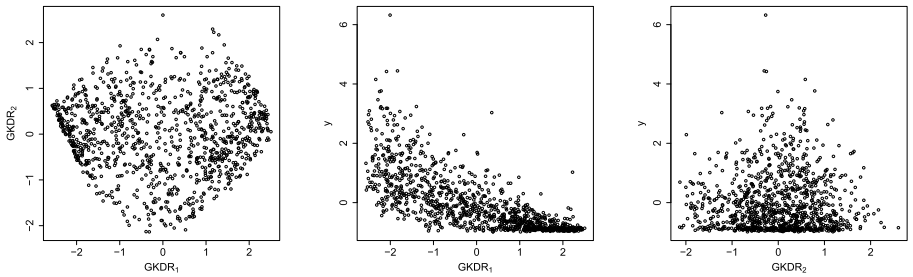


FIG 3. *Two-dimensional sufficient predictors for PM 2.5 data.*

We apply our method to this dataset in a season-by-season fashion. We only report the results for the winter seasons, when the average PM2.5 concentration was the highest across each year. We smooth the original time series data using eight-hour moving windows. We randomly select two-thirds of samples for training and the rest for testing. We standardize both the response and the explanatory variables in the training set so that they have mean 0 and standard deviation 1.

The dimension of the minimal FSDR σ-field is determined as two applying

the heuristic method in Section 4.2. Four knots are used to estimate the transformations. After applying the GKDR method, we obtain two sufficient predictors which are denoted as $GKDR_1$ and $GKDR_2$ respectively and given as follows.

$$\begin{aligned} GKDR_1 =& -0.65\hat{\phi}_1(DEWP) + 0.54\hat{\phi}_2(HUMI) - 0.24\hat{\phi}_3(PRES) \\ & - 0.03\hat{\phi}_4(TEMP) + 0.48\hat{\phi}_5(IWS); \\ GKDR_2 =& 0.56\hat{\phi}_1(DEWP) + 0.55\hat{\phi}_2(HUMI) + 0.45\hat{\phi}_3(PRES) \\ & + 0.19\hat{\phi}_4(TEMP) + 0.38\hat{\phi}_5(IWS). \end{aligned}$$

In the above, $\hat{\phi}_i's$ are the estimated univariate transformations of the explanatory variables for $i = 1, \ldots, 5$. The estimated transformations are depicted in Figure 2. All five transformations are highly nonlinear and have absolute coefficients larger than 0.1 in at least one of the sufficient predictors.

Figure 3(a) shows the scatter plot for $GKDR_2$ versus $GKDR_1$. The boundary constraint observed in this plot can be largely explained by the imbalanced distribution of the explanatory variables. Figures 3(b) and (c) show the marginal scatter plots for Y versus $GKDR_1$ and $GKDR_2$, respectively. From Figure 3(b), a decreasing trend in the mean response can be observed as $GKDR_1$ increases. From Figures 3(b) and (c), we also see a larger variation for Y when $GKDR_1$ takes small values or $GKDR_2$ takes medium values. To further investigate how the sufficient predictors affect the response, we explore the 3-d plot for Y versus $GKDR_1$ and $GKDR_2$; See Figure 4(a). We not only observe a complicated nonlinear relationship between the mean response and the sufficient predictors, but also a dependency of the variance of the response on the sufficient predictors, the latter of which is called the heteroscedasticity phenomenon. We first use the loess method to fit a nonparametric regression model between the response and the two sufficient predictors. To model the heteroscedasticity phenomenon, we use the residual-based estimator proposed by Fan and Yao [8] to estimate the residual standard deviation. In particular, we use the loess method to fit a nonparametric regression model between the squared residuals and the two sufficient predictors, which gives a local smoothing estimator of the residual variance. The estimator of residual standard deviation is derived by taking the squared root of the estimator of residual variance. Figures 4(a) and (b) show the 3-d plot for the estimated regression surface and the estimated residual standard deviation. These plots reveal highly nonlinear relationship for both the mean response and the residual standard deviation versus the sufficient predictors. Figure 4(c) shows the plot for the scaled residual which is defined as the original residual divided by the estimated residual standard deviation, and indicates that heteroscedasticity phenomenon has been removed. Figure 4(d) shows the scatter plot for the observed response versus the fitted response, and their correlation is calculated to be 0.758. We use this fitted model to predict PM2.5 concentration level in the testing set, and the correlation between the observed response and the predicted response is calculated to be 0.766. The above results reveal that the sufficient predictors estimated by the GKDR method are informative and shed new lights on the impacts of the meteorological attributes on the PM2.5

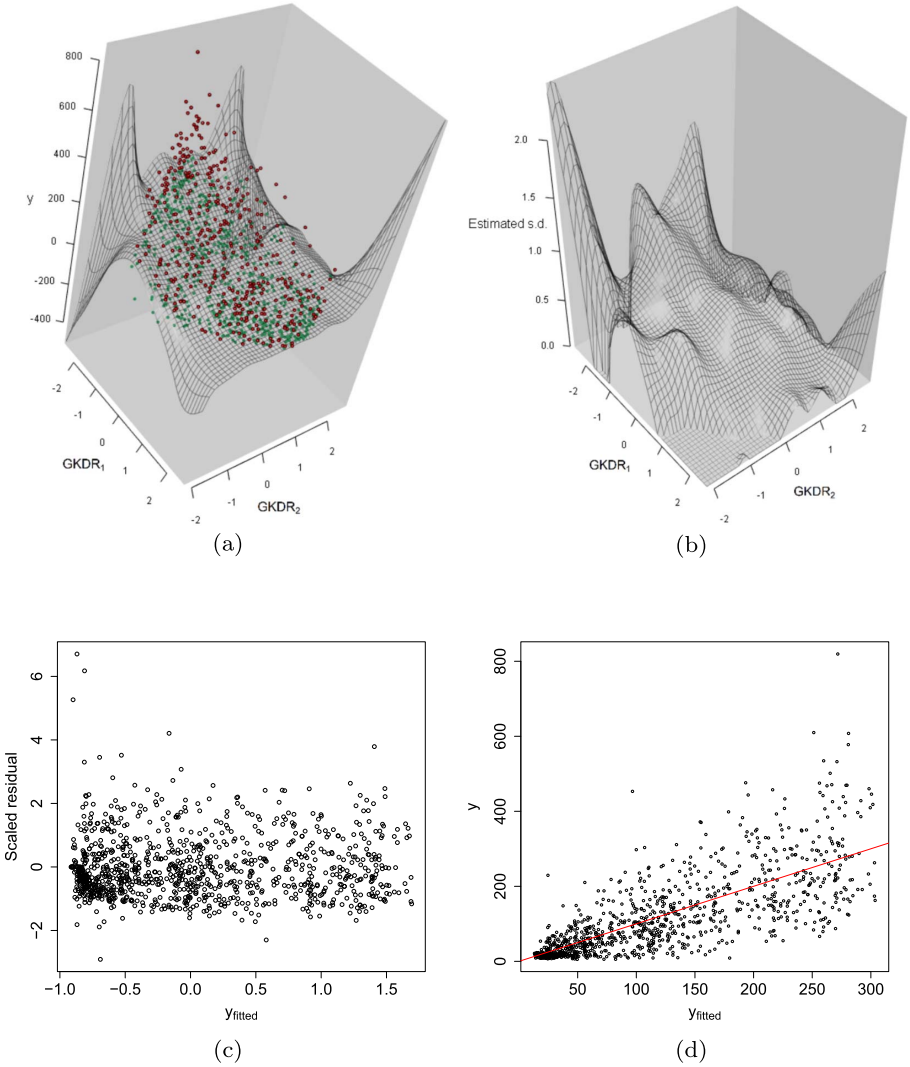


FIG 4. *Nonparametric regression using the sufficient predictors for PM 2.5 data. (a) shows the 3-d plot for the estimated regression surface, where red and green points denote samples lie above and below the surface, respectively. (b) shows the 3-d plot of the estimated residual standard deviation. (c) shows the plot for the scaled residual, which is defined as the original residual divided by the estimated residual standard deviation. (d) shows the scatter plot for the observed response versus the fitted response.*

concentration levels.

PM2.5 refers to the atmospheric particulate matter (PM) with aerodynamic diameter of less than 2.5 micrometers, and is reported to cause respiratory and cardiovascular diseases [13]. It was reported that humidity is highly relevant to

PM2.5 pollution during winter in Beijing. In particular, high level of humidity would lead to abundance in water-soluble components such as inorganic ions, and further concentration of PM2.5 pollution episodes [3]. Our analysis also supports the impact of humidity on PM2.5 concentration. From Figure 2(b) we can see that when humidity is high, the value of first sufficient predictor $GKDR_1$ tends to be lower. This results in a higher expected value of PM2.5 concentration, as shown in Figure 3(b). We also observe some interesting patterns of other meteorological attributes. When temperature is at relatively high or low value, the value of second sufficient predictor $GKDR_2$ tends to be higher or lower, respectively. These result in a lower expected value of PM2.5 concentration, as shown in Figure 3(c). Low PM2.5 concentration in extremely cold weather can be attributed to the strong cold air from the north or northwest with high wind speed [14]. When temperature is at relatively median value and demand for heating supply is high, production intensity of coal-fired power plants is high, which leads to high emission of atmospheric chemical such as SO_2 and NO_2 and further high PM2.5 concentration [28]. Whereas when temperature is relatively high in winter and demand for heating supply is reduced, production intensity of coal-fired power plants decreases, which leads to low emission of atmospheric chemical and further low PM2.5 concentration. This interesting pattern of how the temperature affects the PM2.5 concentration has not been reported by previous studies.

5. Discussion

In this article, we propose to use FSDR σ-fields to characterize flexible SDR and show that the minimal FSDR σ-field σ_{B^*,ϕ^*} exists under some mild conditions. To estimate σ_{B^*,ϕ^*}, we develop the GKDR method and demonstrate its effectiveness through extensive simulation experiments and two real life applications. We believe that the proposed GKDR method can become a useful tool for sufficient dimension reduction in high dimension regression analysis. There are three directions to further improve the current work. First, we have established the consistency result for the GKDR Stage I estimator. However, the large sample properties of the GKDR Stage II estimator is left to be an open problem. Second, in this article, we have proposed a heuristic method to determine the dimension of the minimal FSDR σ-field. More systematic approaches with theoretical justifications will greatly enhance the GKDR method, and thus need to be developed. Third, when the dimension of $\boldsymbol{X}$ increases, the computational intensity of the current GKDR method can also quickly increase. Therefore, more efficient algorithms need to be developed. One approach to mitigating the computational complexity is to incorporate variable selection into the GKDR method. We are currently working on these three directions and hope to report results in the future.

Appendix A: Proofs of the theorems and propositions

A.1. Proof of Proposition 1

Proof. By contradiction, if $\zeta_i(X_i) \not\equiv 0$ or $\psi_i(\boldsymbol{X}_{(-i)}) \not\equiv 0$, then there exist x_i and $\tilde{x}_i$ such that $\mathcal{X}_{(-i|x_i)} \neq \varnothing$, $\mathcal{X}_{(-i|\tilde{x}_i)} \neq \varnothing$, and $\zeta_i(x_i) \neq \zeta_i(\tilde{x}_i)$. Under Assumption 2, we can find $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{X}$ whose i-th entry are $x_i, \tilde{x}_i$ respectively and $\boldsymbol{x}_{(-i)} = \tilde{\boldsymbol{x}}_{(-i)}$. By conditions we have $\zeta_i(x_i) = -\psi_i(\boldsymbol{x}_{(-i)}) = -\psi_i(\tilde{\boldsymbol{x}}_{(-i)}) = \zeta_i(\tilde{x}_i)$, which leads to contradiction. This completes the proof of Proposition 1. □

A.2. Proof of Theorem 2

Proof. We need two useful lemmas in the proof of Theorem 2. The proofs of these two lemmas are provided in Appendix A.2.1 and A.2.2.

Lemma 1. *Under Assumption 2, if* $\sigma_{B^{(2)},\boldsymbol{\phi}^{(2)}} \subseteq \sigma_{B^{(1)},\boldsymbol{\phi}^{(1)}}$, *then* $\sigma\big(\phi_i^{(2)}(X_i)\big) \subseteq \sigma\big(\phi_i^{(1)}(X_i)\big)$ *for* $i \in \mathcal{I}_{B^{(2)}} \cup \mathcal{J}_{B^{(2)}}$, $\mathcal{I}_{B^{(2)}} \subseteq \mathcal{I}_{B^{(1)}}$, $\mathcal{J}_{B^{(2)}} \subseteq \mathcal{I}_{B^{(1)}} \cup \mathcal{J}_{B^{(1)}}$, $\mathcal{K}_{B^{(2)}} \supseteq \mathcal{K}_{B^{(1)}}$, *and* $rank(B^{(2)}) \leq rank(B^{(1)})$.

Lemma 2. *Under Assumption 2, let* $\sigma_{B^{(2)},\boldsymbol{\phi}^{(2)}}$ *and* $\sigma_{B^{(1)},\boldsymbol{\phi}^{(1)}}$ *be two FSDR* σ*-field in* $\mathcal{A}$ *satisfying* $\sigma_{B^{(2)},\boldsymbol{\phi}^{(2)}} \subseteq \sigma_{B^{(1)},\boldsymbol{\phi}^{(1)}}$, *then there exist constant numbers* u_i *and* v_i *such that* $\phi_i^{(2)}(X_i) = u_i * \phi_i^{(1)}(X_i) + v_i$ *for* $i \in \mathcal{J}_{B^{(1)}} \cap \mathcal{J}_{B^{(2)}}$.

Now we return to prove Theorem 2. The main idea is to use Zorn's Lemma. First, we will show that any decreasing FSDR σ-field chain in the set $\mathcal{A}$ is close. In fact, suppose $\sigma_{B^{(1)},\boldsymbol{\phi}^{(1)}} \supseteq \sigma_{B^{(2)},\boldsymbol{\phi}^{(2)}} \supseteq \cdots \supseteq \sigma_{B^{(k)},\boldsymbol{\phi}^{(k)}} \supseteq \ldots$ is a decreasing FSDR σ-field chain in $\mathcal{A}$, without loss of generality, we assume $\sigma\big(\phi_i^{(k+1)}(X_i)\big) \subseteq \sigma\big(\phi_i^{(k)}(X_i)\big)$ using Lemma 1. Let $\mathcal{I}_k$, $\mathcal{J}_k$, and $\mathcal{K}_k$ be the abbreviation for $\mathcal{I}_{B^{(k)}}$, $\mathcal{J}_{B^{(k)}}$, and $\mathcal{K}_{B^{(k)}}$ respectively. By Lemma 1 we have $\mathcal{I}_1 \supseteq \mathcal{I}_2 \supseteq \ldots$, since the cardinal of $\mathcal{I}_1$ is finite, there exists an integer t such that $\mathcal{I}_k = \mathcal{I}_t$ for all $k \geq t$. Consider $k \geq t$, by Lemma 1 again we have $\mathcal{K}_t \subseteq \mathcal{K}_{t+1} \subseteq \ldots \mathcal{K}_k \subseteq \ldots$. Similarly, there exists an integer s such that $\mathcal{K}_k = \mathcal{K}_s$ for all $k \geq s$. Therefore, for $k \geq s$, we have $\mathcal{I}_k = \mathcal{I}_s = \mathcal{I}$, $\mathcal{J}_k = \mathcal{J}_s = \mathcal{J}$, and $\mathcal{K}_k = \mathcal{K}_s = \mathcal{K}$. Without loss of generality, we assume that $s = 1$. For i in $\mathcal{I}$, we have $\sigma\big(\phi_i^{(k+1)}(X_i)\big) \subseteq \sigma\big(\phi_i^{(k)}(X_i)\big)$, so we can find one ϕ_i^* such that $\sigma\big(\phi_i^*(X_i)\big) = \lim_{k\to+\infty}\sigma\big(\phi_i^{(k)}(X_i)\big)$; And by Lemma 2, for j in $\mathcal{J}$, we have $\phi_j^{(k+1)}(X_j) = g_j^{(k)}(\phi_j^{(k)}(X_j))$, where $g_j^{(k)}$ is a linear function. For simplicity we can set $\phi_j^{(k+1)}(X_j) = \phi_j^{(k)}(X_j) = \phi_j^*(X_j)$, where $var[\phi_j^*(X_j)] = 1$. Note that $\sigma\Big(\bigcup_{i\in\mathcal{I}} \phi_i^{(k+1)}(X_i), B_{\mathcal{J}}^{(k+1)T}\boldsymbol{\phi}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}})\Big) = \sigma\big(B^{(k+1)T}\boldsymbol{\phi}^{(k+1)}(\boldsymbol{X})\big) \subseteq \sigma\big(B^{(k)T}\boldsymbol{\phi}^{(k)}(\boldsymbol{X})\big) = \sigma\Big(\bigcup_{i\in\mathcal{I}} \phi_i^{(k)}(X_i), B_{\mathcal{J}}^{(k)T}\boldsymbol{\phi}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}})\Big)$, so we have $\sigma\big(\phi_i^{(k+1)}(X_i)\big) \subseteq \sigma\big(\phi_i^{(k)}(X_i)\big)$ for $i \in \mathcal{I}$ and $\sigma\big(B_{\mathcal{J}}^{(k+1)T}\boldsymbol{\phi}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}})\big) \subseteq \sigma\big(B_{\mathcal{J}}^{(k)T}\boldsymbol{\phi}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}})\big)$. Therefore, we can find one B^* such that $span(B_{\mathcal{J}}^*) = \lim_{k\to+\infty} span(B_{\mathcal{J}}^{(k)})$, and further

$\bigcap_{k=1}^{+\infty}\sigma_{B^{(k)},\boldsymbol{\phi}^{(k)}}=\sigma\Big(\bigcup_{i\in\mathcal{I}}\phi_i^*(X_i),B_{\mathcal{J}}^{*T}\boldsymbol{\phi}_{\mathcal{J}}^*(\boldsymbol{X}_{\mathcal{J}})\Big)=\sigma_{B^*,\boldsymbol{\phi}^*}$. Since the intersection of any two SDR σ-fields is still an SDR σ-field, as stated in the proof of Theorem 1 in [16], we have $Y\perp\!\!\!\perp\boldsymbol{X}|\sigma_{B^*,\boldsymbol{\phi}^*}$. Therefore, any decreasing σ-field chain in the set $\mathcal{A}$ is close. Finally, applying Zorn's Lemma, we can prove that there exists a unique minimal σ-field in $\mathcal{A}$.

Furthermore, suppose $(\tilde{B},\tilde{\boldsymbol{\phi}})$ is another generating pair of $\sigma_{B^*,\boldsymbol{\phi}^*}$. Then we can construct two decreasing σ-field chains, which begin at $\sigma_{\tilde{B},\tilde{\boldsymbol{\phi}}}$ and $\sigma_{B^*,\boldsymbol{\phi}^*}$ and converge to $\sigma_{B^*,\boldsymbol{\phi}^*}$ and $\sigma_{\tilde{B},\tilde{\boldsymbol{\phi}}}$, respectively. Using the properties of decreasing σ-field chains shown previously, we have $span(\tilde{B})=span(B^*)$, $\sigma\big(\tilde{\phi}_i(X_i)\big)=\sigma\big(\phi_i^*(X_i)\big)$ for all $i\in\mathcal{I}_{B^*}$, and there exist constant numbers u_j and v_j such that $\tilde{\phi}_j=u_j*\phi_j^*+v_j$ for all $j\in\mathcal{J}_{B^*}$. This completes the proof of Theorem 2. □

A.2.1. Proof of Lemma 1

Proof. Without loss of generality, we assume both $B^{(1)}$ and $B^{(2)}$ are of full column rank. For $i\in\mathcal{I}_{B^{(2)}}\cup\mathcal{J}_{B^{(2)}}$, we can always find a p-dimensional column vector $\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_p)^T\in span(B^{(2)})$ such that $\lambda_i\neq 0$. Since $\sigma\big(\boldsymbol{\lambda}^T\boldsymbol{\phi}^{(2)}(\boldsymbol{X})\big)\subseteq\sigma_{B^{(2)},\boldsymbol{\phi}^{(2)}}\subseteq\sigma_{B^{(1)},\boldsymbol{\phi}^{(1)}}\subseteq\sigma\big(\boldsymbol{\phi}^{(1)}(\boldsymbol{X})\big)$, we have

$$\lambda_1\phi_1^{(2)}(X_1)+\cdots+\lambda_p\phi_p^{(2)}(X_p)=f\big(\phi_1^{(1)}(X_1),\ldots,\phi_p^{(1)}(X_p)\big),\tag{35}$$

where f is an unknown function. For $x_i\neq\tilde{x}_i$, if $\phi_i^{(1)}(x_i)=\phi_i^{(1)}(\tilde{x}_i)$, $\mathcal{X}_{(-i)|x_i}\neq\varnothing$, and $\mathcal{X}_{(-i)|\tilde{x}_i}\neq\varnothing$, by Assumption 2 we can find $\boldsymbol{x}=(x_1,\ldots,x_i,\ldots,x_p)^T\in\mathcal{X}$ and $\tilde{\boldsymbol{x}}=(x_1,\ldots,\tilde{x}_i,\ldots,x_p)^T\in\mathcal{X}$, then from (35) we have $\phi_i^{(2)}(x_i)=\phi_i^{(2)}(\tilde{x}_i)$. Therefore, $\sigma\big(\phi_i^{(2)}(X_i)\big)\subseteq\sigma\big(\phi_i^{(1)}(X_i)\big)$.

Next, we will show that if $i\in\mathcal{J}_{B^{(2)}}$, then $i\in\mathcal{I}_{B^{(1)}}\cup\mathcal{J}_{B^{(1)}}$. By contradiction, if $i\in\mathcal{K}_{B^{(1)}}$, then there exists $\boldsymbol{\beta}=(b_1,\ldots,b_p)^T\in span(B^{(2)})$ such that $b_i\neq 0$ and $\sigma\big(\boldsymbol{\beta}^T\boldsymbol{\phi}^{(2)}(\boldsymbol{X})\big)\subseteq\sigma_{B^{(2)},\boldsymbol{\phi}^{(2)}}\subseteq\sigma_{B^{(1)},\boldsymbol{\phi}^{(1)}}\subseteq\sigma\big(\boldsymbol{\phi}_{(-i)}^{(1)}(\boldsymbol{X}_{(-i)})\big)$, so $\sum_{k=1}^p b_k\phi_k^{(2)}(X_k)=g\big(\boldsymbol{\phi}_{(-i)}^{(1)}(\boldsymbol{X}_{(-i)})\big)$, where g is an unknown function. Since $\sigma\big(\phi_k^{(2)}(X_k)\big)\subseteq\sigma\big(\phi_k^{(1)}(X_k)\big)$ for $k\in\mathcal{I}_{B^{(2)}}\cup\mathcal{J}_{B^{(2)}}$ and $b_k=0$ for $k\in\mathcal{K}_{B^{(2)}}$, so $b_i\phi_i^{(2)}(X_i)=h\big(\boldsymbol{\phi}_{(-i)}^{(1)}(\boldsymbol{X}_{(-i)})\big)$, where h is an unknown function. By Proposition 1, we have $b_i\phi_i^{(2)}(X_i)=h\big(\boldsymbol{\phi}_{(-i)}^{(1)}(\boldsymbol{X}_{(-i)})\big)=0$, which is a contradiction. Therefore we have $\mathcal{J}_{B^{(2)}}\subseteq\mathcal{I}_{B^{(1)}}\cup\mathcal{J}_{B^{(1)}}$. The proof of $\mathcal{I}_{B^{(2)}}\subseteq\mathcal{I}_{B^{(1)}}$ is similar, and we immediately have $\mathcal{K}_{B^{(2)}}\supseteq\mathcal{K}_{B^{(1)}}$ after the property that $\mathcal{I}_{B^{(s)}}$, $\mathcal{J}_{B^{(s)}}$, and $\mathcal{K}_{B^{(s)}}$ form a partition of indices $\{1,\ldots,p\}$ for $s=1,2$. Note that

$$\sigma\big(B^{(1)T}\boldsymbol{\phi}^{(1)}(\boldsymbol{X})\big)=\sigma\big(\bigcup_{i\in\mathcal{I}_{B^{(2)}}}\phi_i^{(1)}(X_i),\bigcup_{i\in\mathcal{I}_{B^{(1)}}\setminus I_{B^{(2)}}}\phi_i^{(1)}(X_i),B_{\mathcal{J}}^{(1)T}\boldsymbol{\phi}_{\mathcal{J}_{B^{(1)}}}^{(1)}(\boldsymbol{X}_{\mathcal{J}_{B^{(1)}}})\big),$$
$$\sigma\big(B^{(2)T}\boldsymbol{\phi}^{(2)}(\boldsymbol{X})\big)=\sigma\big(\bigcup_{i\in\mathcal{I}_{B^{(2)}}}\phi_i^{(2)}(X_2),B_{\mathcal{J}}^{(2)T}\boldsymbol{\phi}_{\mathcal{J}_{B^{(2)}}}^{(2)}(\boldsymbol{X}_{\mathcal{J}_{B^{(2)}}})\big).$$

Therefore we have $rank(B^{(1)})\geq rank(B^{(2)})$ immediately. This completes the proof of Lemma 1. □

A.2.2. Proof of Lemma 2

Proof. Without loss of generality, we can assume $\sigma\big(\phi_i^{(2)}(X_i)\big) \subseteq \sigma\big(\phi_i^{(1)}(X_i)\big)$ for all i by Lemma 1. This is because for $i \in \mathcal{K}_{B^{(2)}}$, we can simply set $\phi_i^{(2)}(x_i) = \phi_i^{(1)}(x_i)$ which doesn't change $\sigma_{B^{(2)},\boldsymbol{\phi}^{(2)}}$. In addition, we assume both $B^{(1)}$ and $B^{(2)}$ are of full column rank. For convenience, we introduce the notation that $\boldsymbol{Z} = \boldsymbol{\phi}^{(1)}(\boldsymbol{X}) = (Z_1, \ldots, Z_p)^T$ and $\boldsymbol{\phi}^{(2)}(\boldsymbol{X}) = \boldsymbol{g}(\boldsymbol{Z}) = \big(g_1(Z_1), \ldots, g_p(Z_p)\big)^T$, where $\boldsymbol{g} = (g_1, \ldots, g_p)^T$ are unknown functions, so the condition becomes $\sigma\big(B^{(2)T}\boldsymbol{g}(\boldsymbol{Z})\big) \subseteq \sigma(B^{(1)T}\boldsymbol{Z})$. Then for any column vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)^T \in span(B^{(2)})$, we have

$$\boldsymbol{a}^T\boldsymbol{g}(\boldsymbol{Z}) = a_1g_1(Z_1) + a_2g_2(Z_2) + \cdots + a_pg_p(Z_p) = h(\beta_1^T\boldsymbol{Z}, \beta_2^T\boldsymbol{Z}, \ldots, \beta_d^T\boldsymbol{Z}),$$

where $B^{(1)} = (\beta_1, \beta_2, \ldots, \beta_d)$, $d = rank(B^{(1)})$, and h is an unknown function. If $d = p$, then $\mathcal{I}_{B^{(1)}} = \{1, \ldots, p\}$ and $\mathcal{J}_{B^{(1)}} = \varnothing$, the proof is done. Else, we have $d < p$, and suppose $\boldsymbol{z} = (z_1, \ldots, z_p)^T \in supp_{\boldsymbol{Z}}$ where $supp_{\boldsymbol{Z}}$ denote the support of $\boldsymbol{Z}$. Let $span(B^{(1)})^\perp$ denote the orthogonal complement space of $span(B^{(1)})$. Then for any column vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T \in span(B^{(1)})^\perp$ and $t \in \mathbb{R}$ where $|t| < \delta$ and $\delta > 0$ is sufficiently small, $\boldsymbol{a}^T\boldsymbol{g}(\boldsymbol{z} + t\boldsymbol{\lambda}) = h\big(\beta_1^T(\boldsymbol{z} + t\boldsymbol{\lambda}), \beta_2^T(\boldsymbol{z} + t\boldsymbol{\lambda}), \ldots, \beta_d^T(\boldsymbol{z}+t\boldsymbol{\lambda})\big) = h(\beta_1^T\boldsymbol{z}, \beta_2^T\boldsymbol{z}, \ldots, \beta_d^T\boldsymbol{z})$ is a constant function with respect to t. That is,

$$a_1g_1(z_1 + \lambda_1t) + \cdots + a_pg_p(z_p + \lambda_pt) = constant, \text{ for any } |t| < \delta.$$

Therefore, we have

$$a_1[g_1(z_1 + \lambda_1t) - g_1(z_1)] + \cdots + a_p[g_p(z_p + \lambda_pt) - g_p(z_p)] = 0. \tag{36}$$

Note that (36) holds for all $\boldsymbol{z} \in supp_{\boldsymbol{Z}}$, t sufficiently small such that $\boldsymbol{z} + t\boldsymbol{\lambda} \in supp_{\boldsymbol{Z}}$, $\boldsymbol{a} \in span(B^{(2)})$, and $\boldsymbol{\lambda} \in span(B^{(1)})^\perp$. For any $i \in \mathcal{J}_{B^{(1)}} \cap \mathcal{J}_{B^{(2)}}$, we have $i \notin \mathcal{I}_{B^{(1)}}$ and thus there exists $\boldsymbol{\lambda} \in span(B^{(1)})^\perp$ such that $\lambda_i \neq 0$, and $i \notin \mathcal{K}_{B^{(2)}}$ and thus there exists $\boldsymbol{a} \in span(B^{(2)})$ such that $a_i \neq 0$. Then by (36), $g_i(z_i + \lambda_it) - g_i(z_i)$ is constant function with respect to z_i, which means g_i is a linear function. This completes the proof of Lemma 2. □

A.3. Proof of Theorem 3

Proof. From the proof of Theorem 2 we have $span(B^*) \subseteq span(\tilde{B})$. Since $rank(\tilde{B}) = rank(B^*)$, so $span(B^*) = span(\tilde{B})$, and further we have $\mathcal{I}_{B^*} = \mathcal{I}_{\tilde{B}}$ and $\mathcal{J}_{B^*} = \mathcal{J}_{\tilde{B}}$. Note that $\sigma\big(B^{*T}\boldsymbol{\phi}^*(\boldsymbol{X})\big) = \sigma\Big(\bigcup_{i\in\mathcal{I}_{B^*}} \phi_i^*(X_i), B_{\mathcal{J}}^{*T}\boldsymbol{\phi}_{\mathcal{J}}^*(\boldsymbol{X}_{\mathcal{J}})\Big) \subseteq \sigma\Big(\bigcup_{i\in\mathcal{I}_{\tilde{B}}} \tilde{\phi}_i(X_i), \tilde{B}_{\mathcal{J}}^T\tilde{\boldsymbol{\phi}}_{\mathcal{J}}(\boldsymbol{X}_{\mathcal{J}})\Big) = \sigma\big(\tilde{B}^T\tilde{\boldsymbol{\phi}}(\boldsymbol{X})\big)$, so $\sigma\big(\phi_i^*(X_i)\big) \subseteq \sigma\big(\tilde{\phi}_i(X_i)\big)$ for $i \in \mathcal{I}_{B^*}$. By Lemma 2, there exist constant numbers u_j and v_j such that $\tilde{\phi}_j = u_j * \phi_j^* + v_j$ for $j \in \mathcal{J}_{B^*}$. This completes the proof of Theorem 3. □

A.4. Proof of Theorem 4

Proof. Following the well known Law of Total Variance, we have

$$Var\big[g(Y)|B^T\boldsymbol{\phi}(\boldsymbol{X})\big] = E\Big[Var\big[g(Y)|\boldsymbol{X}\big]|B^T\boldsymbol{\phi}(\boldsymbol{X})\Big] + Var\Big[E\big[g(Y)|\boldsymbol{X}\big]|B^T\boldsymbol{\phi}(\boldsymbol{X})\Big].$$

By taking expectation over $B^T\boldsymbol{\phi}(\boldsymbol{X})$, we have

$$E\Big[Var\big[g(Y)|B^T\boldsymbol{\phi}(\boldsymbol{X})\big]\Big] = E\Big[Var\big[g(Y)|\boldsymbol{X}\big]\Big] + E\Big[Var\Big[E\big[g(Y)|\boldsymbol{X}\big]|B^T\boldsymbol{\phi}(\boldsymbol{X})\Big]\Big].$$

If Assumption 3 holds, by Proposition 3 in [10] we have

$$\begin{aligned}\langle g, \Sigma_{YY|\boldsymbol{X}} g\rangle_{\mathcal{H}_\mathcal{Y}} &= E\Big[Var\big[g(Y)|\boldsymbol{X}\big]\Big],\\ \langle g, \Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}} g\rangle_{\mathcal{H}_\mathcal{Y}} &= E\Big[Var\big[g(Y)|B^T\boldsymbol{\phi}(\boldsymbol{X})\big]\Big],\end{aligned}$$

for all $y \in \mathcal{H}_\mathcal{Y}$. Combining these two equations, we have

$$\big\langle g, (\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}} - \Sigma_{YY|\boldsymbol{X}})g\big\rangle_{\mathcal{H}_\mathcal{Y}} = E\Big[Var\Big[E\big[g(Y)|\boldsymbol{X}\big]|B^T\boldsymbol{\phi}(\boldsymbol{X})\Big]\Big].$$

The term on the right hand side of the last equation is non-negative, therefore $\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}} \geq \Sigma_{YY|\boldsymbol{X}}$, and we have

$$\begin{aligned}\Sigma_{YY|\boldsymbol{X}} = \Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}} &\Leftrightarrow \big\langle g, (\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}} - \Sigma_{YY|\boldsymbol{X}})g\big\rangle_{\mathcal{H}_\mathcal{Y}} \equiv 0\\ &\Leftrightarrow E\Big[Var\Big[E\big[g(Y)|\boldsymbol{X}\big]|B^T\boldsymbol{\phi}(\boldsymbol{X})\Big]\Big] \equiv 0\\ &\Leftrightarrow Var\Big[E\big[g(Y)|\boldsymbol{X}\big]|B^T\boldsymbol{\phi}(\boldsymbol{X})\Big] \equiv 0\\ &\Leftrightarrow E\big[g(Y)|\boldsymbol{X}\big] = T\big(B^T\boldsymbol{\phi}(\boldsymbol{X})\big),\ a.s.\ P_{\boldsymbol{X}}\\ &\Leftrightarrow E\big[g(Y)|\boldsymbol{X}\big] = E\big[g(Y)|B^T\boldsymbol{\phi}(\boldsymbol{X})\big],\ a.s.\ P_{\boldsymbol{X}}.\end{aligned}$$

Since $\mathcal{H}_\mathcal{Y}$ is characteristic, this implies that the conditional probability of Y given $\boldsymbol{X}$ is reduced to that of Y given $B^T\boldsymbol{\phi}(\boldsymbol{X})$. This completes the proof of Theorem 4. □

A.5. Proof of Theorem 5

Proof. Let

$$F_{i,q} = \big\{\phi_i \in S_{i,q} : E_{X_i}|\phi_i|^2 \leq 2\big\}; \tag{37}$$

$$\boldsymbol{F}_q = \big\{(\phi_1, \ldots, \phi_p)^T : \phi_i \in F_{i,q},\ \text{for } i = 1, \ldots, p\big\}. \tag{38}$$

It can be shown that $F_{i,q}$ is a compact set with respect to distance $L^2_{X_i}$, and $\boldsymbol{F}_q$ is a compact set with respect to distance $L^2_{\boldsymbol{X}}$. To prove Theorem 5, we will need the following two lemmas. The proof of Lemma 3 is provided in Appendix A.5.1. Lemma 4 comes from Theorem XII.1 in [7] and its proof is omitted here.

Lemma 3. *Under Assumptions 3-5, for a fixed integer $q > 1$, the functions $Tr[\hat{\Sigma}^{B,\boldsymbol{\phi}(n)}_{YY|\boldsymbol{X}}]$ and $Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}]$ are continuous on $\mathbb{S}^p_d(\mathbb{R}) \times \boldsymbol{F}_q$ and*

$$\sup_{B\in\mathbb{S}^p_d(\mathbb{R}),\boldsymbol{\phi}\in\boldsymbol{F}_q} \left|Tr[\hat{\Sigma}^{B,\boldsymbol{\phi}(n)}_{YY|\boldsymbol{X}}] - Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}]\right| \to 0 \quad (n \to \infty)$$

in probability.

Lemma 4. *Let h be a function on $[0,1]$ which is p_0 times differentiable and the p_0-th derivative satisfies $\left|h^{(p_0)}(x_1) - h^{(p_0)}(x_2)\right| \le c|x_1 - x_2|^\nu$ for some $c > 0$ and $0 < \nu \le 1$. Let S_q denote the function space of splines on $[0,1]$ with order 4 and equally spaced knots. If $p_0 + \nu \le 3$, then there exists a function $h_1 \in S_q$ such that*

$$\sup_{0\le x\le 1} |h_1(x) - h(x)| \le c_0 q^{-p}$$

for some c_0 (c_0 depends on h).

We come back to prove Theorem 5. We first establish the consistency result in the spline function space. Let

$$G_{i,q} = F_{i,q} \cap \tilde{L}^2(P_{X_i}); \tag{39}$$

$$\boldsymbol{G}_q = \left\{(\phi_1,\ldots,\phi_p)^T : \phi_i \in G_{i,q}, \text{ for } i = 1,\ldots,p\right\}. \tag{40}$$

For any fixed integer q, Let $\Theta^{(q)}$ denote the collection of normalized generating pairs defined as follows.

$$\Theta^{(q)} = \underset{(B,\boldsymbol{\phi}):B\in\mathbb{S}^p_{d_0}(\mathbb{R}),\boldsymbol{\phi}\in\boldsymbol{G}_q}{argmin} Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}].$$

We first assert that $D(B^{(q)}, \mathbb{B}^p_{d_0}) \to 0$ and $L^2_{\boldsymbol{X}}(\boldsymbol{\phi}^{(q)}, \Phi_1) \to 0$ as $q \to \infty$, where $(B^{(q)}, \boldsymbol{\phi}^{(q)})$ is any generating pair in $\Theta^{(q)}$. Note that $Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}]$ is continuous with respect to B equipped with distance D and $\boldsymbol{\phi}$ equipped with distance $L^2_{\boldsymbol{X}}$. As q increase, Assumption 7 works for the sufficient condition of Lemma 4, and it follows that the $L^2_{X_i}$ distance between spline function space $S_{i,q}$ and $\tilde{\boldsymbol{\phi}}_i$ decrease to 0. Therefore, for $(B^{(q)}, \boldsymbol{\phi}^{(q)}) \in \Theta^{(q)}$, $Tr[\Sigma^{B^{(q)},\boldsymbol{\phi}^{(q)}}_{YY|\boldsymbol{X}}]$ should decrease to $Tr[\Sigma^{\tilde{B},\tilde{\boldsymbol{\phi}}}_{YY|\boldsymbol{X}}]$. From Assumption 6, we immediately have $D(B^{(q)}, \mathbb{B}^p_{d_0}) \to 0$ and $L^2_{\boldsymbol{X}}(\boldsymbol{\phi}^{(q)}, \Phi_1) \to 0$ as $q \to \infty$. Henceforth, for any open set $\tilde{\mathbb{B}}_1 \supseteq \mathbb{B}^p_{d_0}$ and any open set $\tilde{\Phi}_1 \supseteq \Phi_1$, we can find q sufficiently large such that $B^{(q)} \in \tilde{\mathbb{B}}_1$ and $\boldsymbol{\phi}^{(q)} \in \tilde{\Phi}_1$ for any $(B^{(q)}, \boldsymbol{\phi}^{(q)}) \in \Theta^{(q)}$, and thus $\Theta^{(q)} \subseteq \tilde{\mathbb{B}}_1 \times \tilde{\Phi}_1$.

Since $\hat{\phi}_{n,i}$ satisfies that $\frac{1}{n}\sum_{t=1}^n \hat{\phi}_{n,i}(X_i^{(t)}) = 0$ and $\frac{1}{n}\sum_{t=1}^n \left[\hat{\phi}_{n,i}(X_i^{(t)})\right]^2 = 1$, by Law of Large number, we have $E_{X_i}\left[\hat{\phi}_{n,i}(X_i)\right]^2 \le 2$ as n goes to infinity, and thus $\hat{\phi}_{n,i} \in F_{i,q}, i = 1,\ldots,p$ in probability. By Lemma 3, following straightforwardly by standard arguments establishing the consistency of M-estimators (see for example, Section 5.2 in [21]), one can show that the distance between $(\hat{B}_n, \hat{\boldsymbol{\phi}}_n)$ and $\Theta^{(q)}$ converges to zero as n goes to infinity in probability. Therefore $\lim_{n\to\infty} \Pr\left(\hat{B}_n \in \tilde{\mathbb{B}}_1, \hat{\boldsymbol{\phi}}_n \in \tilde{\Phi}_1\right) = 1$. This completes the proof of Theorem 5. □

A.5.1. Proof of Lemma 3

Proof. The proof follows the proof of Proposition 7 in [10]. In particular, under Assumptions 3 and 4, we have $Tr[\hat{\Sigma}^{B,\boldsymbol{\phi}(n)}_{YY|\boldsymbol{X}}]$ and $Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}]$ are continuous on $\mathbb{S}^p_d(\mathbb{R}) \times \boldsymbol{F}_q$ after Lemmas 12-13 in [10]. To prove the uniform convergence of $\big|Tr[\hat{\Sigma}^{B,\boldsymbol{\phi}(n)}_{YY|\boldsymbol{X}}] - Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}]\big|$, we first note that $\big|Tr[\hat{\Sigma}^{B,\boldsymbol{\phi}(n)}_{YY|\boldsymbol{X}}] - Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}]\big| \leq \big|Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}] - Tr[\Sigma_{YY} - \Sigma^{B,\boldsymbol{\phi}}_{Y\boldsymbol{X}}(\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{XX}} + \epsilon_n I_n)^{-1}\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{X}Y}]\big| + \big|Tr[\hat{\Sigma}^{B,\boldsymbol{\phi}(n)}_{YY|\boldsymbol{X}}] - Tr[\Sigma_{YY} - \Sigma^{B,\boldsymbol{\phi}}_{Y\boldsymbol{X}}(\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{XX}} + \epsilon_n I_n)^{-1}\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{X}Y}]\big|$. Under Assumptions 3 and 4, following Lemma 14 in [10], we have

$$\sup_{B\in\mathbb{S}^p_d(\mathbb{R}),\boldsymbol{\phi}\in\boldsymbol{F}_q} \big|Tr[\Sigma^{B,\boldsymbol{\phi}}_{YY|\boldsymbol{X}}] - Tr[\Sigma_{YY} - \Sigma^{B,\boldsymbol{\phi}}_{Y\boldsymbol{X}}(\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{XX}} + \epsilon_n I_n)^{-1}\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{X}Y}]\big| \to 0 \quad (n \to \infty).$$

Under Assumption 5, following Lemma 10 in [10], we have

$$\sup_{B\in\mathbb{S}^p_d(\mathbb{R}),\boldsymbol{\phi}\in\boldsymbol{F}_q} \big|Tr[\hat{\Sigma}^{B,\boldsymbol{\phi}(n)}_{YY|\boldsymbol{X}}] - Tr[\Sigma_{YY} - \Sigma^{B,\boldsymbol{\phi}}_{Y\boldsymbol{X}}(\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{XX}} + \epsilon_n I_n)^{-1}\Sigma^{B,\boldsymbol{\phi}}_{\boldsymbol{X}Y}]\big| = O_p(\epsilon_n^{-1} n^{-1/2}),$$

as n goes to infinity. Therefore, combining the above two results we immediately prove the uniform convergence. □

A.6. Proof of Theorem 6

Proof. The first part of Theorem 6 is a generalization of Theorem 2.2.1 in [6]. If $\sigma(X) \supseteq \sigma\big(\psi(X)\big) \supseteq \sigma\big(\phi(X)\big)$ and $p\big(X|\phi(X)\big)$ is discrete, then both $p\big(X|\psi(X)\big)$ and $p\big(\psi(X)|\phi(X)\big)$ are discrete and $p\big(X|\phi(X)\big) = p\big(X|\psi(X)\big) * p\big(\psi(X)|\phi(X)\big)$. Therefore,

$$\begin{aligned} E_{X|\phi(X)}\big[\log p(X|\phi(X))\big] &= E_{X|\phi(X)}\big[\log p(X|\psi(X)) + \log p(\psi(X)|\phi(X))\big] \\ &= E_{X|\phi(X)}\big[\log p(X|\psi(X))\big] + E_{\psi(X)|\phi(X)}\big[\log p(\psi(X)|\phi(X))\big]. \end{aligned}$$

Therefore, we have

$$\begin{aligned} H\big[X|\phi(X)\big] &= -E_{\phi(X)}\Big[E_{X|\phi(X)}\big[\log p(X|\phi(X))\big]\Big] \\ &= -E_{\phi(X)}\Big[E_{X|\phi(X)}\big[\log p(X|\psi(X))\big]\Big] \\ &\quad -E_{\phi(X)}\Big[E_{\psi(X)|\phi(X)}\big[\log p(\psi(X)|\phi(X))\big]\Big] \\ &= -E\big[\log p(X|\psi(X))\big] - E\big[\log p(\psi(X)|\phi(X))\big] \\ &= H\big[X|\psi(X)\big] + H\big[\psi(X)|\phi(X)\big]. \end{aligned}$$

This completes the proof of the first part of Theorem 6. Furthermore, by definition of Θ_0 and Θ_1, we have $\sigma(X_i) \supseteq \sigma\big(\tilde{\phi}_i(X_i)\big) \supseteq \sigma\big(\phi^*_i(X_i)\big)$ for $i \in \mathcal{I}_{B^*} \cup \mathcal{J}_{B^*}$. Let $w^*_i = \|B^*_{i,\cdot}\|_2$ and $\tilde{w}_i = \|\tilde{B}_{i,\cdot}\|_2$. It can be verified that $w^*_i = \tilde{w}_i > 0$ and $w^*_i H\big[X_i|\phi^*_i(X_i)\big] \geq \tilde{w}_i H\big[X_i|\tilde{\phi}_i(X_i)\big]$ for $i \in \mathcal{I}_{B^*} \cup \mathcal{J}_{B^*}$, and $w^*_i = \tilde{w}_i = 0$

for $i \in \mathcal{K}_{B^*}$. Therefore, $H_{B^*}\big[\boldsymbol{X}|\boldsymbol{\phi}^*(\boldsymbol{X})\big] \geq H_{\tilde{B}}\big[\boldsymbol{X}|\tilde{\boldsymbol{\phi}}(\boldsymbol{X})\big]$, and the equivalence holds if and only if $H\big[X_i|\phi_i^*(X_i)\big] = H\big[X_i|\tilde{\phi}_i(X_i)\big]$ for all $i \in \mathcal{I}_{B^*} \cup \mathcal{J}_{B^*}$, which means $(\tilde{B}, \tilde{\boldsymbol{\phi}}) \in \Theta_0$. This completes the proof of the second part of Theorem 6. □

Acknowledgments

The authors would like to thank the editor, the associate editor, and the referees for their thoughtful suggestions and comments.

References

[1] Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons. MR0534323

[2] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* **80** 580–598. MR0803258

[3] Cheng, Y., He, K.-b., Du, Z.-y., Zheng, M., Duan, F.-k. and Ma, Y.-l. (2015). Humidity plays an important role in the PM2. 5 pollution in Beijing. *Environmental pollution* **197** 68–75.

[4] Chiaromonte, F. and Cook, R. D. (2002). Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics* **54** 768–795. MR1954046

[5] Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association* **86** 328–332. MR1137117

[6] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons. MR2239987

[7] De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C. and De Boor, C. (1978). *A practical guide to splines* **27**. Springer-Verlag New York. MR0507062

[8] Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** 645–660. MR1665822

[9] Fine, S. and Scheinberg, K. (2001). Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research* **2** 243–264.

[10] Fukumizu, K., Bach, F. R., Jordan, M. I. et al. (2009). Kernel dimension reduction in regression. *The Annals of Statistics* **37** 1871–1905. MR2533474

[11] Fukumizu, K. and Leng, C. (2014). Gradient-Based Kernel Dimension Reduction for Regression. *Journal of the American Statistical Association* **109** 359–370. MR3180569

[12] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models* **43**. CRC press. MR1082147

[13] Huang, K., Zhuang, G., Wang, Q., Fu, J., Lin, Y., Liu, T., Han, L. and Deng, C. (2014). Extreme haze pollution in Beijing during January 2013: chemical characteristics, formation mechanism and role of fog processing. *Atmospheric Chemistry and Physics Discussions* **14** 7517–7556.

[14] Jia, Y., Rahn, K. A., He, K., Wen, T. and Wang, Y. (2008). A novel technique for quantifying the regional component of urban aerosol solely from its sawtooth cycles. *Journal of Geophysical Research: Atmospheres* **113**.

[15] Kobayashi, S. and Nomizu, K. (1963). *Foundations of differential geometry* **1**. Interscience publishers New York. MR0152974

[16] Lee, K.-Y., Li, B., Chiaromonte, F. et al. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics* **41** 221–249. MR3059416

[17] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327. MR1137117

[18] Lian, H. and Wang, Q. (2016). Kernel additive sliced inverse regression. *Statistica Sinica* 527–546. MR3497758

[19] Liang, X., Li, S., Zhang, S., Huang, H. and Chen, S. X. (2016). PM2. 5 data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research: Atmospheres* **121**.

[20] Robert, P. and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *Journal of the Royal Statistical Society* **25** 257-265. MR0440801

[21] Van der Vaart, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press. MR1652247

[22] Wang, L. and Yang, L. (2009). Spline estimation of single-index models. *Statistica Sinica* **19** 765. MR2514187

[23] Wang, T. and Zhu, L. (2018). Flexible dimension reduction in regression. *Statistica Sinica* **28** 1009–1029. MR3791098

[24] Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming* **142** 397–434. MR3127080

[25] Williams, C. K. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems* 682–688.

[26] Wu, H. M. (2008). Kernel Sliced Inverse Regression with Applications to Classification. *Journal of Computational and Graphical Statistics* **17** 590-610. MR2528238

[27] Xia, Y., Tong, H., Li, W. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 363–410. MR1924297

[28] Zhang, R., Wang, G., Guo, S., Zamora, M. L., Ying, Q., Lin, Y., Wang, W., Hu, M. and Wang, Y. (2015). Formation of urban fine particulate matter. *Chemical reviews* **115** 3803–3855.

[29] Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association* **101** 1638–1651. MR2279485