Electronic Journal of Statistics
Vol. 15 (2021) 589–629
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1785

Estimating multi-index models with response-conditional least squares

Timo Klock[1], **Alessandro Lanteri**[2] **and Stefano Vigogna**[3]

[1]*Machine Intelligence Department, Simula Research Laboratory, e-mail:* timo@simula.no

[2]*DEMM, Università degli Studi di Milano and Collegio Carlo Alberto, e-mail:* alessandro.lanteri@unimi.it

[3]*MaLGa Center, DIBRIS, Università degli Studi di Genova, e-mail:* vigogna@dibris.unige.it

Abstract: The multi-index model is a simple yet powerful high-dimensional regression model which circumvents the curse of dimensionality assuming $\mathbb{E}[Y|X] = g(A^\top X)$ for some unknown index space A and link function g. In this paper we introduce a method for the estimation of the index space, and study the propagation error of an index space estimate in the regression of the link function. The proposed method approximates the index space by the span of linear regression slope coefficients computed over level sets of the data. Being based on ordinary least squares, our approach is easy to implement and computationally efficient. We prove a tight concentration bound that shows $N^{-1/2}$-convergence, but also faithfully describes the dependence on the chosen partition of level sets, hence providing guidance on the hyperparameter tuning. The estimator's competitiveness is confirmed by extensive comparisons with state-of-the-art methods, both on synthetic and real data sets. As a second contribution, we establish minimax optimal generalization bounds for k-nearest neighbors and piecewise polynomial regression when trained on samples projected onto any $N^{-1/2}$-consistent estimate of the index space, thus providing complete and provable estimation of the multi-index model.

MSC2020 subject classifications: Primary 62G05; secondary 62G08, 62H99.
Keywords and phrases: Multi-index model, sufficient dimension reduction, nonparametric regression, finite sample bounds.

Received June 2020.

1. Introduction

Many recent advances in the analysis of high-dimensional data are based on the observation that real-world data are inherently structured, and the relationship between the features is often of a lower dimensional nature [1, 6, 30, 31, 41, 42]. A popular model incorporating this assumption is the *multi-index model*, which poses the relation between a predictor $X \in \mathbb{R}^D$ and a response $Y \in \mathbb{R}$ as

$$Y = g(A^\top X) + \zeta, \tag{1}$$

where $A \in \mathbb{R}^{D\times d}$ is an unknown full column rank matrix with $d \ll D$, $g : \mathbb{R}^d \to \mathbb{R}$ is an unknown function, and ζ is a noise term with $\mathbb{E}[\zeta|X] = 0$, independent

of X given $A^\top X$. In the following we refer to g as the *link function* and A as the *index space*, assuming, without loss of generality, that the columns of A are orthonormal [16]. Model (1) asserts that the information required to predict the conditional expectation

$$f(x) := \mathbb{E}[Y|X = x] = g(A^\top x) \tag{2}$$

is encoded in the distribution of $A^\top X$. Therefore, knowing the projection $P := AA^\top$ allows to estimate f in a nonparametric fashion with a number of samples scaling with the intrinsic dimension d, rather than the ambient dimension D.

In this work we derive and analyze a method for estimating f under the model assumption (1) from a given data set $\{(X_i, Y_i) : i = 1, \dots, N\}$, where (X_i, Y_i) are independent copies of (X, Y). In the first step we construct an estimate $\hat{P}$ of the projection P, whereas the second step estimates f, respectively g, by means of classical nonparametric estimators on the projected data set $\{(\hat{P}X_i, Y_i) : i = 1, \dots, N\}$. To construct $\hat{P}$ we first divide the range of Y into J subintervals, and assign each sample X_i to a different level set depending on which interval its response Y_i belongs to. After that, we compute the vector of linear regression coefficients on each level set. We do not use such coefficients to locally estimate f, but rather as an estimate of the average gradient of f on the level set. Our estimate $\hat{P}$ is then simply defined as the orthogonal projector onto the subspace generated by the J vectors of linear coefficients. Since our approach is based on solving localized least squares problems, where localization is enforced by conditioning on the response variable, we call our method *response-conditional least squares* (RCLS). A detailed description of RCLS follows in Section 2.

The proposed method is attractive for practitioners, being computationally efficient and easy to implement, with only one hyperparameter (the number of level sets J) to be specified. An additional advantage is that ordinary least squares can be readily exchanged by variants leveraging priors such as sparsity [34, 50] and further. RCLS is also provable, with strong theoretical guarantees neatly derivable from a few reasonable assumptions. In particular, denoting by $\|\cdot\|_F$ the Frobenius norm, we establish a tight concentration bound

$$\left\|\hat{P} - P\right\|_F \lesssim C(J)\sqrt{\frac{D}{N}}, \tag{3}$$

which disentangles the influence of the sample size N and the parameter J on the performance of our estimator (Corollary 8). Furthermore, we provide finite sample generalization bounds for model (1) accounting for the projection error $\|\hat{P} - P\|$, measured in spectral norm $\|\cdot\|$, in the reduced regression problem. We analyze two popular nonparametric methods, namely k-nearest neighbors regression (kNN) and piecewise polynomial regression, and prove that, for s-Hölder regular functions g, the estimator $\hat{f}$ satisfies the generalization bound (up to logarithmic factors)

$$\mathbb{E}\left|\hat{f}(X) - f(X)\right|^2 \lesssim N^{-\frac{2s}{2s+d}} + \left\|\hat{P} - P\right\|^{\min\{2s,2\}}, \tag{4}$$

where $s \in (0, 1]$ in the case of kNN, and $s \in (0, +\infty)$ in the case of piecewise polynomials (Theorems 11 and 14). The bound (4) shows that optimal estimation rates (in the minimax sense) are achieved by traditional regressors for $d = 1$ and $s > \frac{1}{2}$, or $d \geq 2$ and any $s > 0$, provided that $\|\hat{P} - P\| \lesssim N^{-1/2}$. In particular, combining (3) and (4) we obtain that RCLS paired with piecewise polynomial regression produces an optimal estimation of the multi-index model.

Before providing a general literature review in the next section, we note that this work builts on [27, 28], which analyze estimators based on response-conditional least squares vectors for the single-index model and a nonlinear generalization thereof. While we can resort here to some results developed in [27] for determining the accuracy of local least squares vectors, the analysis requires extension to the multi-index case. Furthermore, in the present paper we derive a regression analysis of the link function, complementing index space estimation by RCLS and beyond.

1.1. Related work

Many methods for estimating the index space have been developed in the statistical literature under the name of *sufficient dimension reduction* [35], where the multi-index model is relaxed to

$$Y \perp\!\!\!\perp X | A^\top X. \tag{5}$$

Note that this setting generalizes our problem since (1) and $\zeta \perp\!\!\!\perp X | A^\top X$ imply (5). A space $\mathrm{Im}(A)$ satisfying (5) is called a *dimension reduction subspace*, and if the intersection of such spaces satisfies (5) it is called *central subspace*. Except for degenerate cases, a unique central subspace exists [10, 11]. One can also consider a model where (5) is replaced by $Y \perp\!\!\!\perp \mathbb{E}[Y|X] | A^\top X$, which leads to the definition of *central mean subspace* [13]. In the case of model (1) with $\zeta \perp\!\!\!\perp X | A^\top X$, the space $\mathrm{Im}(A)$ is both the central subspace and the central mean subspace [13]. Thus, we will treat related research under the same umbrella.

The methods for sufficient dimension reduction can broadly be grouped into *inverse regression based methods* and *nonparametric methods* [1, 45]. The first group reverses the regression dependency between X and Y and uses moments of the conditional predictor $X|Y$ to construct a matrix Λ with $\mathrm{Im}(\Lambda) \subseteq \mathrm{Im}(A)$. The most prominent representatives are sliced inverse regression (SIR/SIRII) [38, 39], sliced average variance estimation (SAVE) [12], and contour regression/directional regression (CR/DR) [36, 37] (see Table 1 for the corresponding definition of Λ). Linear combinations of related matrices Λ have been called *hybrid methods* [59]. Furthermore, in the case where X follows a normal distribution, two popular methods are principal Hessian directions (pHd) [40] and iterative Hessian transformations (iHt) [13]. In this setting, Λ is the averaged Hessian matrix of the regression function, which can be efficiently computed using Stein's Lemma.

If $\mathrm{Im}(\Lambda) \subseteq \mathrm{Im}(A)$, eigenvectors corresponding to nonzero eigenvalues of Λ yield an unbiased subspace of the index space $\mathrm{Im}(A)$. A typical assumption to

guarantee this is the *linear conditional mean* (LCM), given by $\mathbb{E}[X|PX] = PX$. It holds, for example, for all elliptically symmetric distributions [38, 45]. Methods based on second order moments usually need in addition the *constant conditional variance* assumption (CCV), which requires $\mathrm{Cov}\,(X|PX)$ to be non-random. In particular, the normal distribution satisfies both LCM and CCV. If $\mathrm{Im}(\Lambda) = \mathrm{Im}(A)$, a method is called *exhaustive*. A condition to ensure exhaustiveness is $\mathbb{E}[v^\top Z|Y]$ being non-degenerate (*i.e.* not almost surely equal to a constant) for all nonzero $v \in \mathrm{Im}(A)$, where Z is the standardization of X. In Table 1 we denote this condition by RCP (random conditional projection), and by RCP^2 when $\mathbb{E}[v^\top Z|Y]$ is replaced by $\mathbb{E}[(v^\top Z)^2|Y]$.

TABLE 1

A summary of prominent inverse regression based methods (plus pHd). We let Z be the standardized X, and (Z', Y') an independent copy of (Z, Y). The table omits details on contour regression [37] (strongly related to DR), iterative Hessian transformations [13] (related to pHd), and hybrid approaches [59] (linear combinations of methods above).

Method		Matrix Λ	$\mathrm{Im}(\Lambda) \subseteq \mathrm{Im}(A)$	$\mathrm{Im}(\Lambda) = \mathrm{Im}(A)$
SIR	[38]	$\mathrm{Cov}\,(\mathbb{E}[Z\|Y])$	LCM	RCP
SIRII	[39]	$\mathbb{E}(\mathrm{Cov}\,(Z\|Y) - \mathbb{E}\,\mathrm{Cov}\,(Z\|Y))^2$	LCM and CCV	N/A
SAVE	[12]	$\mathbb{E}(\mathsf{Id} - \mathrm{Cov}\,(Z\|Y))^2$	LCM and CCV	RCP or RCP^2
DR	[36]	$\mathbb{E}(2\mathsf{Id} - \mathrm{Cov}\,(Z - Z'\|Y, Y'))^2$	LCM and CCV	RCP or RCP^2
pHd	[40]	$\mathbb{E}(Y - \mathbb{E}Y)(X - \mathbb{E}X)(X - \mathbb{E}X)^\top$	normal X	N/A

As inverse regression based methods require only computation of finite sample means and covariances, they are efficient and easy to implement. The matrix Λ is usually estimated by partitioning the range $\mathrm{Im}(Y) = \cup_{\ell=1}^{J} \mathcal{R}_{J,\ell}$ and approximating statistics of $X|Y$ by empirical statistics of $X|Y \in \mathcal{R}_{J,\ell}$. Therefore, only a single hyperparameter, the number of subsets J, needs to be tuned. A strategy for choosing J optimally is not known [45].

Nonparametric methods try to estimate the gradient field of the regression function f based on the observation that the d leading eigenvectors of $\mathbb{E}[\nabla f(X) \nabla f(X)^\top]$ (assuming f is differentiable) span the index space. The concrete implementation of this idea differs between methods. Popular examples are minimum average variance estimation (MAVE), outer product of gradient estimation (OPG), and variants thereof [58]. While MAVE converges to the index space under mild assumptions, it suffers from the curse of dimensionality due to nonparametric estimation of gradients of f. The inverse MAVE (IMAVE) [58] combines MAVE with inverse regression, achieving $N^{-1/2}$-consistency under LCM. Sliced regression [55] collects local MAVE estimates on response slices, producing $N^{-1/2}$-consistent index estimates free of LCM for $d \leq 3$. Furthermore, iterative generalizations of the average derivative estimation (ADE) [21] have been proved to be $N^{-1/2}$-consistent for $d \leq 3$ and $d \leq 4$ [15, 23].

Compared to inverse regression methods, nonparametric methods rely on less stringent assumptions, but are computationally more demanding, require more hyperparameter tuning, and are often more complex to analyze. The relation between inverse regression and nonparametric methods has been investigated in [44, 46] by introducing semiparametric methods. The authors showed that the computational efficiency and simplicity of inverse regression methods come at the cost of assumptions such as LCM/CCV. Moreover, they demonstrated that inverse regression methods can be modified by including a nonparametric estimation step to achieve theoretical guarantees even when LCM/CCV do not hold.

The work presented above focuses mainly on index space estimation, not providing ways to estimate the link function or generalization bounds for the projected regression problem. Other methods have been studied, addressing both dimensionality reduction and regression in the case $d = 1$ [8, 24, 28, 32, 33, 43, 49] or $d \geq 1$ [9, 18, 57]. The multi-index problem was also considered in an active sampling setting, where the user is allowed to query data points (X, Y) and the goal is to minimize the number of queries [16, 22]. Moreover, model (2) has strong ties with shallow neural network models $f(x) = \sum_{i=1}^{m} g_i(a_i^\top x)$, which are currently actively investigated [17, 25, 48, 52].

1.2. Organization of the paper

Section 2 describes RCLS for index space estimation. Section 3 presents theoretical guarantees on the population level and in the finite sample regime. Section 4 establishes the generalization bound (4). Section 5 compares RCLS with state-of-the-art methods on synthetic and real data sets.

1.3. General notation

We let $\mathbb{N}_0$ be the set of natural numbers including 0 and $[m] := \{1, \dots, m\}$ for any positive integer m. We write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. Throughout the paper, C stands for a universal constant that may change on each appearance. We use $\|\cdot\|$ for the Euclidean norm of vectors, and $\|\cdot\|$, $\|\cdot\|_F$ for the spectral and Frobenius matrix norms, respectively. The notation $\mathcal{O}(G(T))$ defines the class of functions $F(T)$ for which there exist C and T_0 such that $|F(T)| \leq CG(T)$ for all $T \geq T_0$. For a symmetric real matrix $A \in \mathbb{R}^{D \times D}$, we denote the ordered eigenvalues as $\lambda_1(A) \geq \cdots \geq \lambda_D(A)$ and the corresponding eigenvectors as $u_1(A), \dots, u_D(A)$. The Moore–Penrose inverse of A is denoted by $A^\dagger$.

We denote expectation and covariance of a random vector X by $\mathbb{E}[X]$ and $\mathrm{Cov}\,(X)$, respectively, and let $\tilde{X} := X - \mathbb{E}[X]$. The sub-Gaussian norm of a random variable Z is $\|Z\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(Z^2/t^2) \leq 2\}$. Similarly, the sub-Exponential norm is $\|Z\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}\exp(|Z|/t) \leq 2\}$. Finally, we abbreviate the mean squared error of an estimator $\hat{f}$ of f by $\mathrm{MSE}(\hat{f}, f) := \mathbb{E}|\hat{f}(X) - f(X)|^2$.

2. Index space estimation by response-conditional least squares

In this section we describe response-conditional least squares (RCLS), first on the population level and then in its actual empirical implementation. Furthermore, we highlight advantages and disadvantages of the approach compared to other methods in the literature (see Section 1.1).

RCLS Let $\mathrm{Im}(Y) = \cup_{\ell=1}^{J} \mathcal{R}_{J,\ell}$ be an arbitrary decomposition of the range into J intervals. For instance, in the case of a bounded range $\mathrm{Im}(Y) = [0, 1)$, we can think of $\mathcal{R}_{J,\ell} := [\frac{\ell-1}{J}, \frac{\ell}{J})$. The vector of linear slope coefficients on the level set $\mathcal{R}_{J,\ell}$ is defined by

$$b_{J,\ell} := \Sigma_{J,\ell}^{\dagger} \, \mathrm{K}_{J,\ell},$$

where

$$\Sigma_{J,\ell} := \mathrm{Cov}\left(X | Y \in \mathcal{R}_{J,l}\right), \quad \text{and} \quad \mathrm{K}_{J,\ell} := \mathrm{Cov}\left(X, Y | Y \in \mathcal{R}_{J,\ell}\right).$$

Intuitively speaking, $b_{J,\ell}$ can be seen as an averaged gradient of the regression function f over the level set $f^{-1}(\mathcal{R}_{J,\ell})$. Taking into account the model (2), we expect $b_{J,\ell}$ to lie in the index space $\mathrm{Im}(A)$, under suitable assumptions. This motivates to approximate the index space by the leading eigenvectors of a (weighted) outer product matrix of the vectors $\{b_{J,\ell} : \ell \in [J]\}$. We thus define the projection

$$P_J := \sum_{i=1}^{d} u_i(M_J) u_i(M_J)^{\top},$$

where

$$M_J := \sum_{\ell=1}^{J} \rho_{J,\ell} b_{J,\ell} b_{J,\ell}^{\top}, \quad \text{and} \quad \rho_{J,\ell} := \mathbb{P}(Y \in \mathcal{R}_{J,\ell}).$$

In practice we have to replace the quantities just defined with sample estimates. To this purpose, we assign the samples $\{(X_i, Y_i) : i \in [N]\}$ to the subsets

$$\mathcal{Y}_{J,\ell} := \{Y_i : Y_i \in \mathcal{R}_{J,\ell}\}, \quad \text{and} \quad \mathcal{X}_{J,\ell} := \{X_i : Y_i \in \mathcal{R}_{J,\ell}\}, \tag{6}$$

which we refer to as level sets in the following. On each level set, we solve the ordinary least squares problem

$$\hat{b}_{J,\ell} := \hat{\Sigma}_{J,\ell}^{\dagger} \hat{\mathrm{K}}_{J,\ell},$$

where

$$\hat{\Sigma}_{J,\ell} := \hat{\mathbb{E}}_{\mathcal{X}_j}(X - \hat{\mathbb{E}}_{\mathcal{X}_{J,\ell}} X)(X - \hat{\mathbb{E}}_{\mathcal{X}_{J,\ell}} X)^{\top},$$
$$\hat{\mathrm{K}}_{J,\ell} := \hat{\mathbb{E}}_{(\mathcal{X}_{J,\ell}, \mathcal{Y}_{J,\ell})}(X - \hat{\mathbb{E}}_{\mathcal{X}_{J,\ell}} X)(Y - \hat{\mathbb{E}}_{\mathcal{Y}_{J\ell}} Y)$$

and $\hat{\mathbb{E}}_{\mathcal{X}_{J,\ell}}$, $\hat{\mathbb{E}}_{\mathcal{Y}_{J,\ell}}$ and $\hat{\mathbb{E}}_{(\mathcal{X}_{J,\ell},\mathcal{Y}_{J,\ell})}$ denote the usual finite sample means. We therefore compute

$$\hat{P}_J(\tilde{d}) := \sum_{i=1}^{\tilde{d}} u_i(\hat{M}_J) u_i(\hat{M}_J)^\top,$$

where

$$\hat{M}_J := \sum_{\ell=1}^{J} \hat{\rho}_{J,\ell} \hat{b}_{J,\ell} \hat{b}_{J,\ell}^\top, \quad \text{and} \quad \hat{\rho}_{J,\ell} := \frac{|\mathcal{X}_{J,\ell}|}{N}.$$

The parameter $\tilde{d} \leq d$ is user-specified and ideally equals $\dim(\operatorname{span}\{b_{J,\ell} : \ell \in [J]\})$ in the limit $N \to \infty$. If this value is unknown, we select it via model selection techniques or by inspecting the spectrum of $\hat{M}_J$. The procedure for the index space estimation using RCLS is summarized in Algorithm 1.

Algorithm 1 Index space estimation via RCLS

Input: Data set $\{(X_i, Y_i) : i \in [N]\}$, parameters J and $\tilde{d}$
Output: Orthoprojector $\hat{P}_J(\tilde{d})$
split data into $\{\mathcal{X}_{J,\ell} : \ell \in [J]\}$ and $\{\mathcal{Y}_{J,\ell} : \ell \in [J]\}$ according to (6)
for $\ell = 1, \ldots, J$ **do**
$\hat{\Sigma}_{J,\ell} := \hat{\mathbb{E}}_{\mathcal{X}_j}(X - \hat{\mathbb{E}}_{\mathcal{X}_{J,\ell}} X)(X - \hat{\mathbb{E}}_{\mathcal{X}_{J,\ell}} X)^\top$
$\hat{\mathrm{K}}_{J,\ell} := \hat{\mathbb{E}}_{(\mathcal{X}_{J,\ell},\mathcal{Y}_{J,\ell})}(X - \hat{\mathbb{E}}_{\mathcal{X}_{J,\ell}} X)(Y - \hat{\mathbb{E}}_{\mathcal{Y}_{J\ell}} Y)$
$\hat{b}_{J,\ell} := \hat{\Sigma}_{J,\ell}^\dagger \hat{\mathrm{K}}_{J,\ell}$
$\hat{\rho}_{J,\ell} := |\mathcal{X}_{J,\ell}|\, N^{-1}$
end for
$\hat{M}_J := \sum_{\ell=1}^{J} \hat{\rho}_{J,\ell} \hat{b}_{J,\ell} \hat{b}_{J,\ell}^\top$
$\hat{P}_J(\tilde{d}) := \sum_{i=1}^{\tilde{d}} u_i(\hat{M}_J) u_i(\hat{M}_J)^\top$

Remark 1 (Choice of partition)**.** The proposed method offers much flexibility in choosing a decomposition of the range $\operatorname{Im}(Y)$ because both practically and theoretically we require fairly minimal assumptions (*e.g.* no need for disjoint or bounded sets $\mathcal{R}_{J,\ell}$). The most stringent requirements in theory are $\dim(M_J) = d$, so that we can exhaustively estimate the index space $\operatorname{Im}(A)$, and a lower bound for the minimal probability mass $\rho_{J,\min} := \min_{\ell \in [J]} \mathbb{P}(Y \in \mathcal{R}_{J,\ell})$, because our bounds suggest that the accuracy degrades linearly in $1/\rho_{J,\min}$.

In practice and without a priori knowledge, we recommend one of the two following choices:

1. using J equisized intervals of $[\min_{i\in[N]} Y_i, \max_{i\in[N]} Y_i]$;
2. using statistically equivalent blocks, *i.e.* split $[\min_{i\in[N]} Y_i, \max_{i\in[N]} Y_i]$ into J connected intervals with equally many samples (± 1 sample).

Option 2 is the statistically robust choice because it balances the mass in level sets and thus generically balances estimation errors incurred due to $\hat{b}_{J,\ell} \approx b_{J,\ell}$ in each level set. However, Option 1 can perform better if the corresponding least squares vectors $\{b_{J,\ell} : \ell \in [J]\}$ are less coherent, leading to a stabler extraction of the leading d-dimensional eigenspace of $\hat{M}_J$.

The influence of parameter J on the accuracy of RCLS, even under the two generic decomposition choices mentioned above, is highly nontrivial, implying that choosing J a priori without problem knowledge is challenging. Practically, we therefore recommend using cross-validation over a range $\{d, d+1, \ldots, J_{\max}\}$, or a subset thereof. If the problem at hand is favorable to the RCLS estimator, choosing large J can lead to an improved accuracy of $\hat{P}_J(\tilde{d})$ because finer decompositions play favorably with estimating the vector $b_{J,\ell}$ in each level set. On the other hand, increasing J also reduces the signal-to-noise ratio in each level set (in the presence of noise ζ) so that large J generically degrades the accuracy of RCLS. We return to the influence of J on the performance of RCLS in Section 3.2.3 and in experiments in Section 5.

Remark 2 (Algorithmic complexity)**.** The main computational demand is constructing the vectors $\{\hat{b}_{J,\ell} : \ell \in [J]\}$. Assuming we use a partition of disjoint level sets $\mathcal{R}_{J,\ell}$, i.e. each sample is only used once in the construction of $\hat{M}_J$, the cost for this is $\mathcal{O}(\sum_{\ell=1}^{J} |\mathcal{X}_{J,\ell}|\, D^2) = \mathcal{O}(ND^2)$.

Comparison of RCLS with inverse regression methods In RCLS, response conditioning serves to localize and produce multiple estimates rather than induce anisotropy in the marginal distribution (*e.g.* no conditioning is required in the single-index case); hence, it is not a typical inverse regression method. At the same time, it shares the same general advantages: it is simple, computationally efficient, and provable. On par with all inverse regression methods, RCLS requires the LCM assumption. Although it is often more or at least as accurate as second order inverse regression methods, such as CR and DR, it does not need the CCV assumption. This is a major generalization since, as pointed out in [45], assuming both LCM and CCV for all directions reduces X to the normal distribution. RCLS low computational cost matches that of typical inverse regression estimates (except CR, which is $\mathcal{O}(N^2D^2)$).

Comparison of RCLS with nonparametric methods Essentially relying on gradient field estimation, RCLS has strong ties with nonparametric methods, but it has lower computation cost and it is easier to implement. Note that nonparametric methods typically involve kernel smoothing, leading to complexities quadratic in the sample size N. Such costs are linearizable resorting for example to nearest neighbor truncation, but while naive kNN still requires the computation of $\mathcal{O}(N^2)$ distances, hierarchical structures for fast neighbor search, such as k-d and cover trees [3, 4], imply constants exponential in the dimension D, not to mention the overhead resulting from cross-validating the number of neighbors. Cross-validation is in principle also required for bandwidth selection, even for joint tuning of two different bandwidths [55], since optimal choices beyond rules of thumb (*e.g.* the "normal reference") are to date an open problem. Last but not least, kernel estimates are sensitive to the curse of dimensionality, whose overcoming requires further complications, algorithmic tweaks, initializations and iterative procedures [55, 58].

3. Guarantees for RCLS

All quantities thus far are defined through the random vector (X, Y) without using the regression function. In fact, in this section we can technically avoid specifying the regression function by defining P as the orthogonal projection onto the minimal dimensional subspace such that

(A1) $Y \perp\!\!\!\perp X|PX$.

As mentioned in Section 1.1, (A1) uniquely defines $\mathrm{Im}(P)$ except for degenerate cases, which we exclude here. Moreover, (1) with $\zeta \perp\!\!\!\perp X|A^\top X$ implies (A1).

In the following analysis, we also require the following assumptions.

(A2) $\mathbb{E}[X|PX] = PX$ almost surely;
(A3) X and Y are sub-Gaussian random variables.

(A2) is the LCM assumption introduced in Section 1.1 and is required in all inverse regression based techniques like SIR, SAVE or DR. It is satisfied for example for any elliptical distribution and ensures $\mathrm{Im}(M_J) \subseteq \mathrm{Im}(A)$ as shown in Proposition 3 below. (A3) is maximally general to use the tools developed in the framework of sub-Gaussian random variables, namely finite sample concentration bounds. Examples of sub-Gaussian random variables include bounded distributions, the normal distribution, or more generally random variables for which all one-dimensional marginals have tails that exhibit a Gaussian-like decay after a certain threshold [54].

3.1. Population level

The population level results are summarized in the following proposition.

Proposition 3. *If (X, Y) satisfies (A1) and (A2), then $b_{J,\ell} \in \mathrm{Im}(A)$ for any $\ell \in [J]$ and any J. Also, $\mathrm{Im}(M_J) \subseteq \mathrm{Im}(A)$, with equality iff $\lambda_d(M_J) > 0$.*

We need the following result for the proof of Proposition 3.

Lemma 4. *Let $Q := \mathsf{Id} - P$. Under (A1) and (A2), we get*

(a) $\mathbb{E}[X|Y] = \mathbb{E}[PX|Y]$ *almost surely, or equivalently* $\mathbb{E}[QX|Y] = 0$*;*
(b) $\mathrm{Cov}(X \mid Y) = \mathrm{Cov}(PX \mid Y) + \mathrm{Cov}(QX \mid Y)$ *almost surely.*

Proof. (a). The towering property of conditional expectations yields $\mathbb{E}[X|Y] = \mathbb{E}[\mathbb{E}[X|PX, Y]|Y]$. Assumption (A1) implies $Y \perp\!\!\!\perp X|PX$, and thus $\mathbb{E}[X|PX, Y] = \mathbb{E}[X|PX] = PX$ by assumption (A2).
(b). By the law of total covariance,

$$\begin{aligned}\mathrm{Cov}(PX, QX|Y) &= \mathbb{E}[\mathrm{Cov}(PX, QX|PX, Y)\,|Y] \\ &\quad + \mathrm{Cov}(\mathbb{E}[QX|PX, Y], \mathbb{E}[PX|PX, Y]|Y) = 0 + 0,\end{aligned}$$

where we used $\mathbb{E}[X|PX, Y] = PX$ as shown in the proof of (a). Therefore,

$$\mathrm{Cov}(X \mid Y) = \mathrm{Cov}(PX + QX \mid Y) = \mathrm{Cov}(PX \mid Y) + \mathrm{Cov}(QX \mid Y). \qquad \square$$

Proof of Proposition 3. We only show that $b_{J,\ell} \in \mathrm{Im}(A)$ for all $R_{J,\ell}$, since $\mathrm{Im}(M_J) \subseteq \mathrm{Im}(A)$ follows immediately. We have

$$\begin{aligned}\mathrm{Cov}\left(QX, Y|Y \in \mathcal{R}_{J,\ell}\right) &= \mathbb{E}\left[\mathrm{Cov}\left(QX, Y|Y\right)|Y \in \mathcal{R}_{J,\ell}\right] \\ &\quad + \mathrm{Cov}\left(\mathbb{E}[QX|Y], \mathbb{E}[Y|Y]|Y \in \mathcal{R}_{J,\ell}\right) \\ &= 0 + \mathrm{Cov}\left(\mathbb{E}[QX|Y], Y|Y \in \mathcal{R}_{J,\ell}\right) = 0,\end{aligned}$$

where the last equality follows from $\mathbb{E}[QX|Y] = 0$ by Lemma 4. Therefore

$$\begin{aligned}\mathrm{K}_{J,\ell} &= \mathrm{Cov}\left(PX, Y|Y \in \mathcal{R}_{J,\ell}\right) + \mathrm{Cov}\left(QX, Y|Y \in \mathcal{R}_{J,\ell}\right) \\ &= \mathrm{Cov}\left(PX, Y|Y \in \mathcal{R}_{J,\ell}\right) + 0.\end{aligned}$$

Furthermore, statement (b) of Lemma 4 implies

$$\Sigma_{J,\ell} = \mathrm{Cov}\left(PX|Y \in \mathcal{R}_{J,\ell}\right) + \mathrm{Cov}\left(QX|Y \in \mathcal{R}_{J,\ell}\right),$$

hence the eigenspace of $\Sigma_{J,\ell}$ decomposes orthogonally into eigenspaces of $\mathrm{Cov}\left(PX|Y \in \mathcal{R}_{J,\ell}\right)$ and of $\mathrm{Cov}\left(QX|Y \in \mathcal{R}_{J,\ell}\right)$. The same holds for $\Sigma_{J,\ell}^{\dagger}$ because the eigenvectors are precisely the same as for $\Sigma_{J,\ell}$. This implies $\Sigma_{J,\ell}^{\dagger} z \in \mathrm{Im}(P)$ for all $z \in \mathrm{Im}(P)$, and the result follows by

$$b_{J,\ell} = \Sigma_{J,\ell}^{\dagger} \mathrm{K}_{J,\ell} = \Sigma_{J,\ell}^{\dagger} \mathrm{Cov}\left(PX, Y|Y \in \mathcal{R}_{J,\ell}\right) \in A. \qquad \square$$

Exhaustiveness Proposition 3 ensures exhaustiveness of RCLS (on the population level) whenever d out of the J least squares vectors $b_{J,\ell}$ are linearly independent. Even when this is not the case, we believe that RCLS generically finds a subspace of the index space that accounts for most of the variability in f, thereby allowing for a *sufficient* dimension reduction. The rationale behind this is that the $b_{J,\ell}$'s can be interpreted as averaged gradients over approximate level sets, and thus they provide samples of the first order behavior of f along the chosen partition. This claim is supported numerically in Section 5.2, where RCLS performs better or as good as all inverse regression based methods listed in Table 1.

Analyzing the exhaustiveness of inverse regression estimators is challenging since in general it is easy to construct examples where some directions of the index space only show up in the tails of (X, Y). This also justifies why most typical exhaustiveness conditions such as RCP and RCP^2 are formulated on the nonquantized level, and therefore do not quite imply exhaustiveness of the actual quantized estimator. The only exception we are aware of is [37, Theorem 3.1], where sufficient conditions for the exhaustiveness of the estimator are provided by decoupling the roles of regression function and noise.

Lastly, we mention that it is possible to further enrich the space $\mathrm{Im}(M_J)$ by adding outer products of vectors $Bb_{J,\ell}$ for matrices B which map $\mathrm{Im}(P)$ to $\mathrm{Im}(P)$. This resembles the idea behind the iHt method [14], where B is chosen as a positive power of the average residual- or response-based Hessian matrix [14, 40]. Other choices are powers of $\Sigma_{J,\ell}$ or $\Sigma_{J,\ell}^{\dagger}$, which map $\mathrm{Im}(P)$ to $\mathrm{Im}(P)$ under (A1) and (A2).

3.2. Finite sample regime

We now analyze the finite sample performance of $\hat{P}_J(\tilde{d})$ as an estimator for the orthoprojector P_J. Our main result (Corollary 8) establishes convergence at rate $N^{-1/2}$, while additionally tracking the dependence on the hyperparameter J. Convergence will be proved using the Davis–Kahan theorem followed by concentration inequalities. For hyperparameter characterization, it will be crucial to exploit anisotropy using the techniques recently derived in [27].

To understand the role of anisotropy in RCLS, recall that P_J is the projection onto the span of least squares vectors $\{b_{J,\ell} : \ell \in [J]\}$, which are obtained after conditioning responses in range subintervals, namely $Y \in \mathcal{R}_{J,\ell}$. Response conditioning induces anisotropy in the distribution of predictors, as encoded in the spectral properties of the conditional covariance matrix $\Sigma_{J,\ell} = \mathrm{Cov}\,(X|Y \in \mathcal{R}_{J,l})$. When concentrating on $\Sigma_{J,\ell}$, an isotropic bound would pay maximum variance factors indiscriminately for all directions. An anisotropic bound, on the other hand, separates the directions of smaller variance, allowing to capture directional dependencies induced by the conditioning and hence, ultimately, by the hyperparameter J.

We expound on anisotropic concentrations for ordinary least squares in Section 3.2.1. After these technical preliminaries, we bound the estimation error of RCLS in Section 3.2.2. In Section 3.2.3 we show numerically that our bounds accurately describe the influence of the hyperparameter. Throughout all sections we let $X|_{J,\ell}$ and $Y|_{J,\ell}$ denote the random variables X and Y conditioned on $Y \in \mathcal{R}_{J,\ell}$. By (A3) and Lemma 18, $X|_{J,\ell}$ and $Y|_{J,\ell}$ are sub-Gaussian whenever $\rho_{J,\ell} > 0$, which implies that $\|X|_{J,\ell}\|_{\psi_2}$ and $\|Y|_{J,\ell}\|_{\psi_2}$ are finite. Moreover, we define $P_{J,\ell}$ as the orthoprojector onto $\mathrm{span}\{b_{J,\ell}\}$ and $Q_{J,\ell} := \mathsf{Id} - P_{J,\ell}$.

3.2.1. Anisotropic concentrations

An anisotropic concentration bound for $\hat{b}_{J,\ell} - b_{J,\ell}$ uses the orthogonal decomposition

$$\hat{b}_{J,\ell} - b_{J,\ell} = P_{J,\ell}(\hat{b}_{J,\ell} - b_{J,\ell}) + Q_{J,\ell}(\hat{b}_{J,\ell} - b_{J,\ell}) = P_{J,\ell}(\hat{b}_{J,\ell} - b_{J,\ell}) + Q_{J,\ell}\hat{b}_{J,\ell},$$

and finds separate bounds for the terms $P_{J,\ell}(\hat{b}_{J,\ell} - b_{J,\ell})$ and $Q_{J,\ell}\hat{b}_{J,\ell}$. To see why those terms play different roles when estimating $\mathrm{Im}(P)$, let us consider the illustrative case of the single-index model, where $P = aa^\top$ for some $a \in \mathbb{S}^{D-1}$. We can estimate a by the direction of any $b_{J,\ell}$, because any nonzero $b_{J,\ell}$ is aligned with a under (A1) and (A2). Using few algebraic manipulations we have, with $Q := \mathsf{Id} - P$,

$$\left\| \frac{\hat{b}_{J,\ell}}{\left\|\hat{b}_{J,\ell}\right\|} - a \right\| = \left\| \frac{\hat{b}_{J,\ell}}{\left\|\hat{b}_{J,\ell}\right\|} - \frac{b_{J,\ell}}{\|b_{J,\ell}\|} \right\| \leq \frac{\left\|Q\hat{b}_{J,\ell}\right\|}{\|b_{J,\ell}\| - \left\|P(\hat{b}_{J,\ell} - b_{J,\ell})\right\|}, \tag{7}$$

whenever $\hat{b}_{J,\ell}^\top b_{J,\ell} > 0$. This reveals that the error is dominated by $\|Q\hat{b}_{J,\ell}\|$, whereas $\|P(\hat{b}_{J,\ell} - b_{J,\ell})\|$ is a higher order error term as soon as $|\mathcal{X}_{J,\ell}|$ is sufficiently

large. A similar observation will be established for higher dimensional index spaces in (11) below.

Anisotropic concentration bounds for ordinary least squares vectors have been recently provided in [27]. To restate the bounds, we introduce a directional sub-Gaussian condition number $\kappa_{J,\ell} := \kappa(P_{J,\ell}, X|_{J,\ell}) \vee \kappa(Q_{J,\ell}, X|_{J,\ell})$, where (recall $\tilde{Z} = Z - \mathbb{E}Z$)

$$\kappa(L, X) := \left\| L \operatorname{Cov}(X)^{\dagger} \tilde{X} \right\|_{\psi_2}^2 \left\| L\tilde{X} \right\|_{\psi_2}^2 .$$

As described in [27], $\kappa(L, X)$ is related to the restricted matrix condition number defined by $\tilde{\kappa}(L, \operatorname{Cov}(X)) := \|L \operatorname{Cov}(X)^{\dagger} L\| \, \|L \operatorname{Cov}(X) L\|$, which measures the heterogeneity of eigenvalues of $\operatorname{Cov}(X)$, when restricting the eigenspaces to $\operatorname{Im}(L)$. In fact, if X follows a normal distribution, the sub-Gaussian norm is a tight variance proxy and $\kappa(L, X)$ differs from $\tilde{\kappa}(L, \operatorname{Cov}(X))$ by a constant factor that only depends on the precise definition of the sub-Gaussian norm.

We further introduce the standardized random variable $\tilde{Z}|_{J,\ell} := \Sigma_{J,\ell}^{-1/2} \tilde{X}|_{J,\ell}$, where $\Sigma_{J,\ell}^{-1/2}$ is the matrix square root of $\Sigma_{J,\ell}^{\dagger}$. As a consequence of the standardization, we have $\operatorname{Cov}(\tilde{Z}|_{J,\ell}) = \mathsf{Id}_D$.

Lemma 5 (Anisotropic ordinary least squares bounds)**.** *Let $J \in \mathbb{N}$, $\ell \in [J]$ and assume (A3). For fixed $u > 0$, $\varepsilon > 0$, with $|\mathcal{X}_{J,\ell}| > C(D + u)(\|\tilde{Z}|_{J,\ell}\|_{\psi_2}^4 \vee \varepsilon^{-2})$, we get, with probability at least $1 - \exp(-u)$,*

$$\left\| P_{J,\ell}(b_{J,\ell} - \hat{b}_{J,\ell}) \right\| \leq \varepsilon \sqrt{\kappa_{J,\ell}} \left\| \tilde{Y}|_{J,\ell} \right\|_{\psi_2} \left\| P_{J,\ell} \Sigma_{J,\ell}^{\dagger} \tilde{X}|_{J,\ell} \right\|_{\psi_2}, \tag{8}$$

$$\left\| Q_{J,\ell} \hat{b}_{J,\ell} \right\| \leq \varepsilon \sqrt{\kappa_{J,\ell}} \left\| \tilde{Y}|_{J,\ell} \right\|_{\psi_2} \left\| Q_{J,\ell} \Sigma_{J,\ell}^{\dagger} \tilde{X}|_{J,\ell} \right\|_{\psi_2} . \tag{9}$$

Furthermore, we have

$$\|b_{J,\ell}\| \leq 2 \left\| \tilde{Y}|_{J,\ell} \right\|_{\psi_2} \left\| P_{J,\ell} \Sigma_{J,\ell}^{\dagger} \tilde{X}|_{J,\ell} \right\|_{\psi_2} . \tag{10}$$

Proof. The concentration bounds are precisely [27, Lemma 14] adjusted to the notation used here. (10) follows from $b_{J,\ell} = P_{J,\ell} \Sigma_{J,\ell}^{\dagger} \operatorname{Cov}(X|_{J,\ell}, Y|_{J,\ell})$ and [27, Lemma 6.5]. □

Equations (8) and (9) in Lemma 5 reveal that the concentration of $P_{J,\ell}(b_{J,\ell} - \hat{b}_{J,\ell})$ and $Q_{J,\ell}\hat{b}_{J,\ell}$ scale with sub-Gaussian norms of $P_{J,\ell}\Sigma_{J,\ell}^{\dagger}\tilde{X}$, and $Q_{J,\ell}\Sigma_{J,\ell}^{\dagger}\tilde{X}$ (we can intuitively think of $\|P_{J,\ell}\Sigma_{J,\ell}^{\dagger}P_{J,\ell}\|$, and $\|Q_{J,\ell}\Sigma_{J,\ell}^{\dagger}Q_{J,\ell}\|$). In many scenarios, both norms, if viewed as functions of the parameter J, behave very differently. This is because increasing the number of level sets J typically reduces the variance in the direction of the least squares solution $b_{J,\ell}$, and therefore increases $\|P_{J,\ell}\Sigma_{J,\ell}^{\dagger}P_{J,\ell}\|$, while $\|Q_{J,\ell}\Sigma_{J,\ell}^{\dagger}Q_{J,\ell}\|$ is often not affected. The effect is particularly strong for single-index models with monotone link functions, as illustrated in Figure 1, but it can also be observed in more general scenarios, for instance if f follows a monotone single-index model locally on one of the level

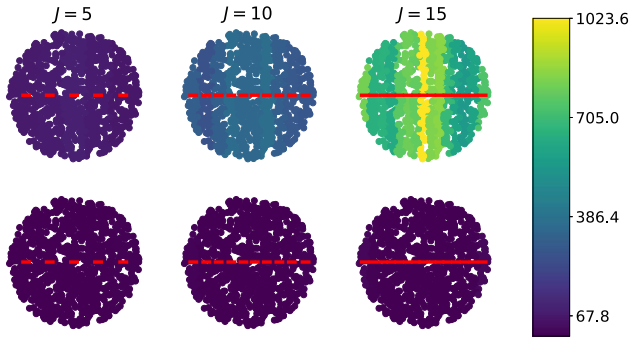


FIG 1. *We sample* $X \sim Uni(\{X : \|X\| \leq 1\})$ *and* $Y = 1/(1+\exp(-X_1))$*, and show* $\|P\Sigma^{\dagger}_{J,\ell}P\|$ *(top) and* $\|Q\Sigma^{\dagger}_{J,\ell}Q\|$ *(bottom) row for* $P = e_1 e_1^{\top}$*,* $Q = \mathsf{Id} - P$*, and all* $\ell \in [J]$ *for* $J \in \{5, 10, 15\}$*. The red lines mark local ordinary least squares vectors. We see that* $\|P\Sigma^{\dagger}_{J,\ell}P\|$ *increases substantially when increasing the number of level sets* J*, while* $\|Q\Sigma^{\dagger}_{J,\ell}Q\|$ *remains roughly constant.*

sets. Recalling (7), using anisotropic concentration is therefore necessary, if we aim at an accurate description of the projection error in terms of both, N and J.

To simplify notation in the following, we introduce the shorthands

$$\eta^{\|}_{J,\ell} := \left\|\tilde{Y}|_{J,\ell}\right\|_{\psi_2} \left\|P_{J,\ell}\Sigma^{\dagger}_{J,\ell}\tilde{X}|_{J,\ell}\right\|_{\psi_2}, \quad \eta^{\perp}_{J,\ell} := \left\|\tilde{Y}|_{J,\ell}\right\|_{\psi_2} \left\|Q_{J,\ell}\Sigma^{\dagger}_{J,\ell}\tilde{X}|_{J,\ell}\right\|_{\psi_2},$$

and $\eta_{J,\ell} := \eta^{\|}_{J,\ell} + \eta^{\perp}_{J,\ell}$.

3.2.2. Concentration bounds for index space estimation

Our goal is now to provide concentration bounds for $\hat{P}_J(\tilde{d})$ around P_J. Using the Davis–Kahan theorem [5, Theorem 7.3.1] we have for $Q_J := \mathsf{Id} - P_J$

$$\left\|\hat{P}_J(\tilde{d})Q_J\right\|_F \leq \frac{\left\|\hat{P}_J(\tilde{d})\left(\hat{M}_J - M_J\right)Q_J\right\|_F}{\lambda_{\tilde{d}}(\hat{M}_J)} \leq \frac{\left\|\left(\hat{M}_J - M_J\right)Q_J\right\|_F}{\lambda_{\tilde{d}}(M_J) - \left\|\hat{M}_J - M_J\right\|}, \tag{11}$$

where we used Weyl's bound [56] to get $\lambda_{\tilde{d}}(\hat{M}_J) \geq \lambda_{\tilde{d}}(M_J) - \|\hat{M}_J - M_J\|$ in the second inequality. It remains to develop concentration bounds for $\|(\hat{M}_J - M_J)Q_J\|_F$, which dictates the projection error, and $\|\hat{M}_J - M_J\|$ to ensure that the denominator does not vanish.

Theorem 6. *Let (A1) - (A3) hold. Fix* $u > 0$*,* $\varepsilon > 0$*, and define*

$$\omega_{J,\max} := \max_{\ell\in[J]} \|\tilde{Z}|_{J,\ell}\|^4_{\psi_2}, \quad \rho_{J,\min} := \min_{\ell\in[J]} \rho_{J,\ell}.$$

Whenever $N > C(D+u+\log(J))(\omega_{J,\max}\ \rho_{J,\min}^{-1} \vee \varepsilon^{-2})$ *we have*

$$\mathbb{P}\left(\left\|\hat{M}_J - M_J\right\|_F \leq \varepsilon \sum_{\ell=1}^{J} \sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}^2\right) \geq 1-\exp(-u). \tag{12}$$

Proof. The first step is to decompose the error according to

$$\begin{aligned}
\left\|M_J - \hat{M}_J\right\|_F &= \left\|\sum_{\ell=1}^{J} \hat{\rho}_{J,\ell}\left(\hat{b}_{J,\ell}\hat{b}_{J,\ell}^\top - b_{J,\ell}b_{J,\ell}^\top\right) - (\rho_{J,\ell}-\hat{\rho}_{J,\ell})\, b_{J,\ell}b_{J,\ell}^\top\right\|_F \\
&\leq \sum_{\ell=1}^{J} \hat{\rho}_{J,\ell}\left\|\hat{b}_{J,\ell}\hat{b}_{J,\ell}^\top - b_{J,\ell}b_{J,\ell}^\top\right\|_F + \sum_{\ell=1}^{J}|\hat{\rho}_{J,\ell}-\rho_{J,\ell}|\left\|b_{J,\ell}b_{J,\ell}^\top\right\|_F \\
&\leq \underbrace{\sum_{\ell=1}^{J}\hat{\rho}_{J,\ell}\left(\left\|\hat{b}_{J,\ell}-b_{J,\ell}\right\| + 2\left\|b_{J,\ell}\right\|\right)\left\|\hat{b}_{J,\ell}-b_{J,\ell}\right\|}_{=:T_1} + \underbrace{\sum_{\ell=1}^{J}|\hat{\rho}_{J,\ell}-\rho_{J,\ell}|\left\|b_{J,\ell}\right\|^2}_{=:T_2}.
\end{aligned}$$

The second term can be bounded using Lemma 5 and 20. Specifically, Lemma 5 implies $\|b_{J,\ell}\| \leq 2\eta_{J,\ell}$, and (28) in Lemma 20 with a union bound argument over $\ell \in [J]$ shows

$$\mathbb{P}\left(\forall \ell \in [J] : |\rho_{J,\ell}-\hat{\rho}_{J,\ell}| \leq \sqrt{\rho_{J,\ell}}\left(\varepsilon \wedge \frac{1}{2}\sqrt{\rho_{J,\ell}}\right)\right) \geq 1-\exp(-u), \tag{13}$$

provided $N > C(u+\log(J))(\rho_{J,\min}^{-1}\vee\varepsilon^{-2})$. Thus we have $T_2 \leq 4\varepsilon\sum_{\ell\in[J]}\sqrt{\rho_{J,\ell}}\eta_{J,\ell}^2$ with probability $1-\exp(-u)$ whenever $N > C(u+\log(J))(\rho_{J,\min}^{-1}\vee\varepsilon^{-2})$.

To bound T_1 we first need to ensure that each level set is sufficiently populated. Using the second case in (13), we have $1/2 \leq \hat{\rho}_{J,\ell}/\rho_{J,\ell} \leq \hat{\rho}_{J,\ell}/\rho_{J,\min}$, and if $N > C\,(D+u+\log(J))\left(\frac{\omega_{J,\max}}{\rho_{J,\min}}\vee\varepsilon^{-2}\right)$ we get

$$|\mathcal{X}_{J,\ell}| = \hat{\rho}_{J,\ell}N > C\,(D+u+\log(J))\left(\|\tilde{Z}|_{J,\ell}\|_{\psi_2}^4 \vee \varepsilon^{-2}\hat{\rho}_{J,\ell}\right). \tag{14}$$

Now we can use Lemma 5 to concentrate $\|\hat{b}_{J,\ell}-b_{J,\ell}\|$, giving the bound

$$\mathbb{P}\left(\forall \ell \in [J] : \left\|\hat{b}_{J,\ell}-b_{J,\ell}\right\| \leq \left(\varepsilon\sqrt{\frac{\kappa_{J,\ell}}{\hat{\rho}_{J,\ell}}}\wedge 1\right)\eta_{J,\ell}\right) \geq 1-\exp(-u). \tag{15}$$

Finally, (15) and $\|b_{J,\ell}\| \leq 2\eta_{J,\ell}$ from Lemma 5 implies

$$\begin{aligned}
T_1 &= \sum_{\ell=1}^{J}\hat{\rho}_{J,\ell}\left(\left\|\hat{b}_{J,\ell}-b_{J,\ell}\right\| + 2\left\|b_{J,\ell}\right\|\right)\left\|\hat{b}_{J,\ell}-b_{J,\ell}\right\| \\
&\leq 5\varepsilon\sum_{\ell=1}^{J}\sqrt{\hat{\rho}_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}^2 \leq 15/2\varepsilon\sum_{\ell=1}^{J}\sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}^2,
\end{aligned}$$

where we used $\hat{\rho}_{J,\ell} \leq 3/2\rho_{J,\ell}$ by (13) in the final step. Taking the union bound over events (13) and (15), all results hold with probability at least $1-2\exp(-u)$ whenever $N > C\,(D+u+\log(J))\,(\frac{\omega_{J,\max}}{\rho_{J,\min}} \vee \varepsilon^{-2})$. The assertion (12) follows with probability at least $1-\exp(-u)$ after adjusting C accordingly. □

Theorem 7. *Let (A1) - (A3) hold. Fix $u > 0$, $\varepsilon > 0$, and define*

$$\omega_{J,\max} := \max_{\ell\in[J]} \|\tilde{Z}|_{J,\ell}\|_{\psi_2}^4, \quad \rho_{J,\min} := \min_{\ell\in[J]} \rho_{J,\ell}.$$

Whenever $N > C(D+u+\log(J))(\omega_{J,\max}\ \rho_{J,\min}^{-1} \vee \varepsilon^{-2})$ we have

$$\mathbb{P}\left(\left\|\left(\hat{M}_J - M_J\right) Q_J\right\|_F \leq \varepsilon \sum_{\ell=1}^{J} \sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}\eta_{J,\ell}^{\perp}\right) \geq 1-\exp(-u).$$

Proof. By the definition of Q_J and $Q_{J,\ell}$, we have $\mathrm{Im}(Q_J) \subseteq \mathrm{Im}(Q_{J,\ell})$. This first allows us to bound

$$\left\|(\hat{M}_J - M_J)Q_J\right\|_F = \left\|\sum_{\ell=1}^{J} \hat{\rho}_{J,\ell}\hat{b}_{J,\ell}\hat{b}_{J,\ell}^{\top}Q_J\right\|_F \leq \sum_{\ell=1}^{J} \hat{\rho}_{J,\ell}\left\|\hat{b}_{J,\ell}\right\|\left\|Q_J\hat{b}_{J,\ell}\right\|$$
$$\leq \sum_{\ell=1}^{J} \hat{\rho}_{J,\ell}\left\|\hat{b}_{J,\ell}\right\|\left\|Q_{J,\ell}\hat{b}_{J,\ell}\right\| \leq \sum_{\ell=1}^{J} \hat{\rho}_{J,\ell}\left(\|b_{J,\ell}\| + \left\|\hat{b}_{J,\ell} - b_{J,\ell}\right\|\right)\left\|Q_{J,\ell}\hat{b}_{J,\ell}\right\|.$$

By the same argument as in the proof of Theorem 6, we have $|\hat{\rho}_{J,\ell} - \rho_{J,\ell}| \leq 1/2\rho_{J,\ell}$ for all $\ell \in [J]$ with probability at least $1-2\exp(-u)$, and thus the number of samples in each level set satisfies (14). Using this together with (8) and (9), and the union bound over $\ell \in [J]$, we get

$$\mathbb{P}\left(\forall \ell \in [J] : \left\|\hat{b}_{J,\ell} - b_{J,\ell}\right\| \leq \eta_{J,\ell}\right) \geq 1-\exp(-u),$$
$$\mathbb{P}\left(\forall \ell \in [J] : \left\|Q_{J,\ell}\hat{b}_{J,\ell}\right\| \leq \varepsilon\sqrt{\frac{\kappa_{J,\ell}}{\hat{\rho}_\ell}}\eta_{J,\ell}^{\perp}\right) \geq 1-\exp(-u).$$

Plugging this, and $\|b_{J,\ell}\| \leq 2\eta_\ell$ by Lemma 5, in the initial decomposition, we get with probability at least $1-4\exp(-u)$

$$\left\|(\hat{M}-M)Q_J\right\|_F \leq \sum_{\ell=1}^{J} \hat{\rho}_{J,\ell}\left(\|b_{J,\ell}\| + \left\|\hat{b}_{J,\ell} - b_{J,\ell}\right\|\right)\left\|Q_{J,\ell}\hat{b}_{J,\ell}\right\|$$
$$\leq 3\varepsilon \sum_{\ell=1}^{J} \sqrt{\hat{\rho}_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}^{\perp}\eta_{J,\ell} \leq \frac{9}{2}\varepsilon \sum_{\ell=1}^{J} \sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}^{\perp}\eta_{J,\ell},$$

where the last step follows from $\hat{\rho}_{J,\ell} \leq 3/2\rho_{J,\ell}$ for all $\ell \in [J]$. By suitable choice of C in the statement, we can absorb the factor $9/2$, and adjust the probability to $1-\exp(-u)$. □

The guarantee for $\hat{P}_J(\tilde{d})$ now follows as a Corollary.

Corollary 8. *Let (A1) - (A3) hold. Fix $u > 0$, $\varepsilon > 0$, and define $\omega_{J,\max} := \max_{\ell\in[J]} \|\tilde{Z}|_{J,\ell}\|_{\psi_2}^4$, $\rho_{J,\min} := \min_{\ell\in[J]} \rho_{J,\ell}$. If $\operatorname{rank}(M_J) = \tilde{d}$ and $\lambda_{\tilde{d}}(M_J) > \gamma_J > 0$ we have with probability at least $1 - \exp(-u)$*

$$\left\|\hat{P}_J(\tilde{d}) - P_J\right\|_F \leq \varepsilon \frac{\sum_{\ell=1}^J \sqrt{\rho_{J,\ell}}\kappa_{J,\ell}\eta_{J,\ell}\eta_{J,\ell}^\perp}{\gamma_J}, \quad \text{whenever}$$

$$N > C(D + u + \log(J)) \left(\frac{\omega_{J,\max}}{\rho_{J,\min}} \vee \left(\frac{\sum_{\ell=1}^J \sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}^2}{\gamma_J} \right)^2 \vee \frac{1}{\varepsilon^2} \right). \tag{16}$$

Proof. Using Weyl's bound $\lambda_{\tilde{d}}(\hat{M}_J) \geq \lambda_{\tilde{d}}(M_J) - \left\|\hat{M}_J - M_J\right\|$ [56], and Theorem 6, we have with probability $1 - \exp(-u)$ the guarantee $\lambda_{\tilde{d}}(\hat{M}_J) \geq \gamma_J - \left\|\hat{M}_J - M_J\right\| > \frac{1}{2}\gamma_J$, whenever the number of samples exceeds

$$N > C(D + u + \log(J)) \left(\frac{\omega_{J,\max}}{\rho_{J,\min}} \vee \left(\frac{\sum_{\ell=1}^J \sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}^2}{\gamma_J} \right)^2 \right)$$

Furthermore, Theorem 7 implies

$$\mathbb{P}\left(\left\| \left(\hat{M}_J - M_J\right) Q_J \right\|_F \leq \varepsilon \sum_{\ell=1}^J \sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}\eta_{J,\ell}^\perp \right) \geq 1 - \exp(-u),$$

whenever $N > C(D + u + \log(J))(\omega_{J,\max}\rho_{J,\min}^{-1} \vee \varepsilon^{-2})$. Using the union bound over both events, the conclusion in the statement follows with probability at least $1 - 2\exp(-u)$ from $\|\hat{P}_J - P_J\|_F \leq \sqrt{2}\|\hat{P}_J(\tilde{d})Q_J\|_F$ (see Lemma 22 in the Appendix), and the Davis–Kahan bound (11). □

Assuming ε^{-2} maximizes (16), Corollary 8 implies the bound

$$\left\|\hat{P}_J(\tilde{d}) - P_J\right\|_F \leq C \frac{\sqrt{1 + \log(J)} \sum_{\ell=1}^J \sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}\eta_{J,\ell}^\perp}{\gamma_J} \sqrt{\frac{D+u}{N}}. \tag{17}$$

It separates the error into a leading factor, which only depends on the hyperparameter J, respectively, the induced level set partition, and a trailing factor, which describes dependencies on $\sqrt{D}$, $N^{-1/2}$ and the confidence parameter u. By using anisotropic bounds from Lemma 5, we obtain a linear dependence on $\eta_{J,\ell}$, which scales like the term $\|\hat{b}_{J,\ell}\|$, and a linear dependence on $\eta_{J,\ell}^\perp$, which scales like $\|Q_J\hat{b}_{J,\ell}\|$. An isotropic concentration bound for $\hat{b}_{J,\ell} - b_{J,\ell}$ would have resulted in $\eta_{J,\ell}^2$, which, judging by Figure 1, leads to a loose characterization of the influence of J onto the error.

Remark 9 (An alternative interpretation of Corollary 8)**.** In certain idealized cases, Corollary 8 allows for an alternative interpretation by choosing J as a function growing with N. To illustrate this, we consider (X, Y) with $\operatorname{Im}(Y) = [0, 1]$ and a partition induced by $\mathcal{R}_{J,\ell} = [\frac{\ell-1}{J}, \frac{\ell}{J})$. Further, we assume there exist universal $C_1, C_2 > 0$ (independent of J) so that

1. we have balanced level sets in the sense that $\rho_{J,\ell} \geq C_1 J^{-1}$,
2. $X_{J,\ell}$ satisfies $\|UX_{J,\ell}\|_{\psi_2}^2 \leq C_2 \|U\Sigma_{J,\ell}U\|_2$ for all matrices U with matching shape (this condition is fairly common, see e.g. [51, Assumption 2.1], and sometimes called strict sub-Gaussianity [2]),
3. partitioning influences local covariances $\Sigma_{J,\ell}$ according to the model

$$\Sigma_{J,\ell} = J^{-2} P_{J,\ell} + Q_{J,\ell}. \tag{18}$$

Condition (18) is an idealization, which asserts that partitioning does not affect $\Sigma_{J,\ell}$ in directions corresponding to $\mathrm{Im}(Q_{J,\ell}) = \mathrm{span}\{b_{J,\ell}\}^\perp$, while reducing the variance in direction $b_{J,\ell}$ by a factor J^{-2}. Using Conditions 1-3, there exists a constant C depending only on C_1, C_2 so that $\omega_{J,\max} \leq C$, $\kappa_{J,\ell} \leq C$, and

$$\eta_{J,\ell}^{\perp} = \|\tilde{Y}|_{J,\ell}\|_{\psi_2} \|Q_{J,\ell}\Sigma_{J,\ell}^{\dagger}\tilde{X}|_{J,\ell}\|_{\psi_2} = \|\tilde{Y}|_{J,\ell}\|_{\psi_2} \sqrt{C_2 \|Q_{J,\ell}\Sigma_{J,\ell}^{\dagger} Q_{J,\ell}\|_2} \leq CJ^{-1},$$
$$\eta_{J,\ell}^{\parallel} = \|\tilde{Y}|_{J,\ell}\|_{\psi_2} \|P_{J,\ell}\Sigma_{J,\ell}^{\dagger}\tilde{X}|_{J,\ell}\|_{\psi_2} \leq \|\tilde{Y}|_{J,\ell}\|_{\psi_2} \sqrt{C_2 \|P_{J,\ell}\Sigma_{J,\ell}^{\dagger} P_{J,\ell}\|_2} \leq C,$$

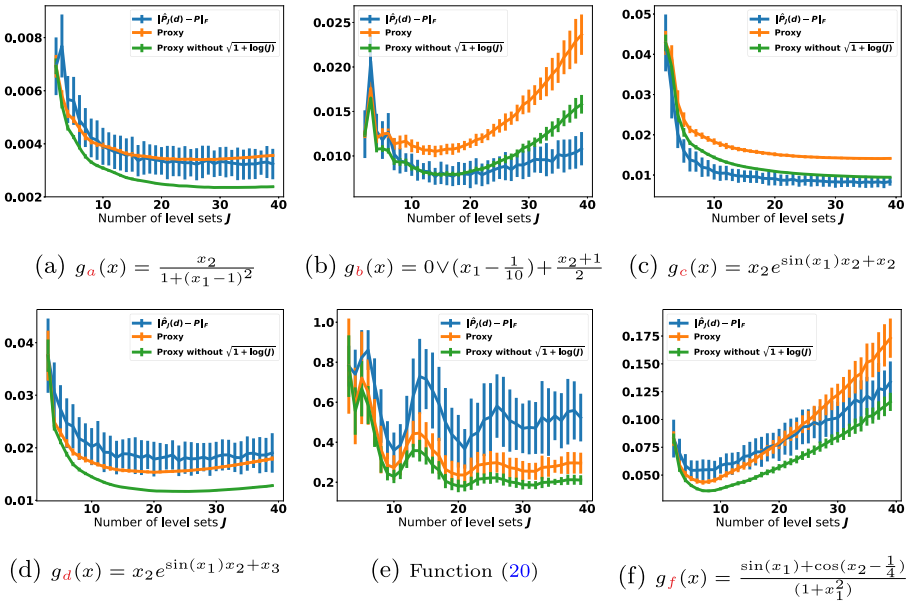
and thus also $\eta_{J,\ell} = \eta_{J,\ell}^{\perp} + \eta_{J,\ell}^{\parallel} \leq C$. Inserting this into Corollary 8 and using $\varepsilon = CJ^{-1/2}$, we obtain with probability at least $1 - \exp(-u)$

$$\left\| \hat{P}_J(\tilde{d}) - P_J \right\|_F \leq \frac{C}{J\gamma_J}, \quad \text{whenever} \quad N > \frac{C(D + u + \log(J))J}{\gamma_J^2}.$$

This reformulation of Corollary 8 suggests to choose J as large as possible so that the condition on N is still satisfied. Assuming $\gamma_J \geq CJ^{-\beta}$ for some $\beta > 0$ uniformly in J, the choice $J \approx \lfloor C(N/(D+u))^{1/(1+2\beta)} \rfloor$ is permitted (neglecting the $\log(J)$-factor for simplicity), which indicates that error rates of up to order $\mathcal{O}(N^{-\frac{1-\beta}{1+2\beta}})$ are possible. If $\beta \in [0, 1/4)$, this improves upon the typical $N^{-1/2}$ rate up to N^{-1} for $\beta = 0$. While Conditions 1-3 and $\gamma_J \geq CJ^{-\beta}$ are unlikely satisfied uniformly for all possible J's, it may hold approximately within a range $J \in \{J_{\min}, J_{\min}+1, \ldots, J_{\max}\}$, resulting in temporary faster error decay. We will return to this phenomenon in our synthetic experiments in Section 5.1, where RCLS achieves faster error decay in certain cases under an optimal choice of J. Indeed, these experiments show the optimal J strongly depends on N, which supports the observations in this remark, but also implies there is no optimal fixed range decomposition for a fixed multi-index problem, independently of N.

3.2.3. Data-driven proxy and tightness of (17)

We now empirically study the tightness of (17) when considering a fixed number of samples N but varying the number of level sets J. First, we develop a data-driven proxy to estimate the leading factor in (17) from a given data set. Afterwards, we compare the proxy with the true error on several synthetic examples.



(a) $g_a(x) = \frac{x_2}{1+(x_1-1)^2}$ (b) $g_b(x) = 0 \vee (x_1 - \frac{1}{10}) + \frac{x_2+1}{2}$ (c) $g_c(x) = x_2 e^{\sin(x_1)x_2 + x_2}$

(d) $g_d(x) = x_2 e^{\sin(x_1)x_2 + x_3}$ (e) Function (20) (f) $g_f(x) = \frac{\sin(x_1)+\cos(x_2-\frac{1}{4})}{(1+x_1^2)}$

FIG 2. *Figures plot the function* $J \mapsto \|\hat{P}_J(d) - P\|_F$ *(blue), and compare it to the right hand side of* (19) *with (orange) and without (green) the union bound factor* $\sqrt{1+\log(J)}$*. The proxy is rescaled by a constant to match* $\|\hat{P}_J(d) - P\|_F$ *for* $J = d$*. Vertical bars indicate the standard deviation.*

Data-driven proxy We have to replace γ_J, $\rho_{J,\ell}$, $\eta_{J,\ell}$, $\eta^{\perp}_{J,\ell}$ and $\kappa_{J,\ell}$ in (17) by quantities that can be estimated from data. The first three quantities are approximated by $\gamma_J \approx \lambda_d(\hat{M}_J)$, $\rho_{J,\ell} \approx \hat{\rho}_{J,\ell}$ and $\eta_{J,\ell} \approx \|\hat{b}_{J,\ell}\|$, where the last replacement is motivated by the fact that $\eta_{J,\ell}$ is used to bound $\|\hat{b}_{J,\ell}\| \leq \|b_{J,\ell}\| + \|b_{J,\ell} - \hat{b}_{J,\ell}\| \lesssim \eta_{J,\ell}$ in the proofs of Theorems 6 and 7. Furthermore, we use the conditional sample covariance $\hat{\Sigma}_{J,\ell}$, and projections $\hat{P}_{J,\ell} := \|\hat{b}_{J,\ell}\|^{-2} \hat{b}_{J,\ell}\hat{b}_{J,\ell}^{\top}$ and $\hat{Q}_{J,\ell} := \mathsf{Id} - \hat{P}_{J,\ell}$, to compute an approximation to $\kappa_{J,\ell}$ by

$$\hat{\kappa}_{J,\ell} := \max\left\{ \left\| \hat{P}_{J,\ell}\hat{\Sigma}_{J,\ell}\hat{P}_{J,\ell} \right\| \left\| \hat{P}_{J,\ell}\hat{\Sigma}^{\dagger}_{J,\ell}\hat{P}_{J,\ell} \right\|, \left\| \hat{Q}_{J,\ell}\hat{\Sigma}_{J,\ell}\hat{Q}_{J,\ell} \right\| \left\| \hat{Q}_{J,\ell}\hat{\Sigma}^{\dagger}_{J,\ell}\hat{Q}_{J,\ell} \right\| \right\}.$$

Note that replacing squared sub-Gaussian norms with spectral norms of the corresponding covariance matrices can underestimate the true value of $\kappa_{J,\ell}$. The same strategy is used for $\eta^{\perp}_{J,\ell}$, *i.e.* we approximate $\eta^{\perp}_{J,\ell}$ by

$$\hat{\eta}^{\perp}_{J,\ell} := \sqrt{\hat{\mathbb{E}}_{\mathcal{Y}_{J,\ell}} \left(Y - \hat{\mathbb{E}}_{\mathcal{Y}_{J,\ell}} Y \right)^2 \left\| \hat{Q}_{J,\ell}\hat{\Sigma}^{\dagger}_{J,\ell}\hat{Q}_{J,\ell} \right\|}.$$

Combining everything, the data-driven proxy for the leading factor in (17) without the union bound factor $\sqrt{1+\log(J)}$ is given by

$$\frac{\sum_{\ell=1}^{J}\sqrt{\rho_{J,\ell}\kappa_{J,\ell}}\eta_{J,\ell}\eta_{J,\ell}^{\perp}}{\gamma_J} \approx \frac{\sum_{\ell=1}^{J}\sqrt{\hat{\rho}_{J,\ell}\hat{\kappa}_{J,\ell}}\left\|\hat{b}_{J,\ell}\right\|\hat{\eta}_{J,\ell}^{\perp}}{\lambda_d(\hat{M}_J)}. \tag{19}$$

In order to reduce the variance in estimating (19), we further restrict the sum to level sets with at least $|\mathcal{X}_{J,\ell}| > 5D$ samples in the experiments below.

Experiments We sample $N = 80000$ points from $\mathrm{Uni}(\{X : \|X\| \leq 1\})$ in $D = 20$ dimensions and set $Y = g(A^\top X) + \zeta$, where $\zeta \sim \mathcal{N}(0, 10^{-4}\mathrm{Var}(f(X)))$ and $A = [e_1|\dots|e_d]$ with e_i being the i-th standard basis vector. Each experiment is repeated 50 times and we report averaged results plus standard deviations for different link functions in Figures 2a-2f. With the only exception of g_e, which is given by

$$\begin{aligned} g_e(x_1,x_2,x_3) &= \sum_{i=1}^{3} h_i(x_i) \quad \text{with} \\ h_1(x_1) &= 1(x_1<0)0.2x_1 + 1(x_1\geq 0)x_1, \\ h_2(x_2) &= 1(x_2<0)0.25x_2 + 1(x_2\geq 0)1.25x_2, \\ h_3(x_3) &= 1(x_3<0)0.2x_3 + 1(x_3\geq 0)1.5x_3, \end{aligned} \tag{20}$$

the link functions we use in our experiments are defined below the respective plot in Figure 2.

We observe that the map $J \mapsto \|\hat{P}_J(d) - P\|_F$ initially decreases when increasing the number of level sets J beyond d, and then either stalls, such as in Figures 2a, 2c, 2d and 2e or increases as in Figures 2b and 2f. This behavior is captured well by the data driven proxy (19). Furthermore, even if the relation $J \mapsto \|\hat{P}_J(d) - P\|_F$ shows several bumps and the index space error depends strongly on J as in 2e, the derived data-driven proxy qualitatively reproduces the same behavior. The experiments suggest that Corollary 8 characterizes the influence of J and the induced level set partition on the projection error well. Furthermore, they raise the question whether J, which minimizes the data-driven proxy (19), can be used for hyperparameter tuning in practice. This is an interesting direction for future work, because choosing J for the related class of inverse regression based methods has been identified as a notoriously difficult problem, for which no good strategies exist [45].

4. Regression in the reduced space

In this section we return to the multi-index model $Y = g(A^\top X) + \zeta$ with $\mathbb{E}[\zeta|X] = 0$ almost surely. Assumption $\zeta \perp\!\!\!\perp X|A^\top X$ is not strictly required in this part. The second step to estimate the model is to learn the link function g, while leveraging the approximated projection $\hat{P} \approx P$, *e.g.* constructed by using RCLS. We restrict our analysis to two popular and commonly used regressors,

namely kNN-regression and piecewise polynomial regression. Our analysis reveals how the error $\|\hat{P} - P\|$ affects kNN and piecewise polynomials, if they are trained on perturbed data $\{(\hat{P}X_i, Y_i) : i \in [N]\}$ instead of $\{(PX_i, Y_i) : i \in [N]\}$. For simplicity, we assume $\hat{P}$ is deterministic and thus statistically independent of $\{(X_i, Y_i) : i \in [N]\}$. In practice, statistical independence can be ensured by using separate data sets for learning $\hat{P}$ and performing the subsequent regression task, for example randomly splitting the data in half.

To study regression rates, smoothness properties of the link function play an important role. We use to the following standard definition [19].

Definition 10. Let $f : \mathbb{R}^D \to \mathbb{R}$, $s_1 \in \mathbb{N}_0$, $s_2 \in (0, 1]$ and $s = s_1 + s_2$. We say f is (L, s)-smooth if partial derivatives $\partial^\alpha f$ exist for all $\alpha \in \mathbb{N}_0^D$ with $\sum_i \alpha_i \leq s_1$, and for all s with $\sum_i \alpha_i = s_1$ we have

$$|\partial^\alpha f(z) - \partial^\alpha(z')| \leq L \|z - z'\|^{s_2}.$$

The minimax rate for nonparametric estimation in $\mathbb{R}^d$ is well known [19, 53] and reads, for (L, s)-smooth regression function f,

$$\mathrm{MSE}\left(\hat{f}, f\right) := \mathbb{E}\left|\hat{f}(X) - f(X)\right|^2 \asymp N^{-\frac{2s}{2s+d}}. \tag{21}$$

Similarly, the rate is a lower bound for nonparametric estimation of the multi-index model with $\dim(P) = d$, because we are still left with a nonparametric regression problem in $\mathbb{R}^d$ once P is identified. In the following, we provide conditions on $\|\hat{P} - P\|$ so that the optimal rate (21) is achieved, when training on perturbed data. In the analysis, we assume that X is sub-Gaussian, $|f(X)| \leq 1$ almost surely, and $\mathrm{Var}(\zeta|X) \leq \sigma_\zeta^2$ almost surely.

4.1. kNN-regression

Let x be a new data point and denote a reordering of the indices by $1(x), \ldots, N(x)$ so that

$$\left\|\hat{P}\left(x - X_{i(x)}\right)\right\| \leq \left\|\hat{P}\left(x - X_{j(x)}\right)\right\| \text{ for all } j \geq i \text{ and all } i,$$

i.e. $i(x)$ is the i-th nearest neighbor to x after projecting onto $\mathrm{Im}(\hat{P})$. The kNN-estimator is defined by $\hat{f}_k(x) := k^{-1} \sum_{i=1}^k Y_{i(x)}$ and the following theorem characterizes the influence of the projection error on the generalization performance. The proof resembles [19, 29] and is given in Appendix A.3.

Theorem 11. *Let g be (L, s)-smooth for $s \in (0, 1]$, and $d > 2s$. For $k = C_k N^{2s/(2s+d)}$, we obtain*

$$MSE\left(\hat{f}_k, f\right) \leq C_1 N^{-\frac{2s}{2s+d}} + C_2 \log(N) \left\|\hat{P} - P\right\|^{2s}, \tag{22}$$

where C_1 depends on $d, \sigma_\zeta, \|X\|_{\psi_2}, C_k, L, s$, and C_2 additionally linearly on $D\|X\|_{\psi_2}^2$.

Remark 12 ($d > 2s$ assumption)**.** The condition $d > 2s$ in Theorem 11 is not due to the error $\|\hat{P} - P\|$, but arises from [29, Lemma 1], where ordinary kNN is analyzed for unbounded marginal distributions. It has been shown in [19] that achieving similar rates for $d \leq 2s$ requires an extra assumption of the marginal distribution of X (boundedness does not suffice).

Remark 13 (Rate optimality)**.** Assuming $\|\hat{P} - P\| \in \mathcal{O}(N^{-1/2})$, we observe that the second term in (22) has order N^{-s}. Therefore, Theorem 11 ensures, up to the logarithmic factor, the optimal rate $N^{-2s/(2s+d)}$ for $d \geq 2$. The logarithmic factor disappears, if the marginal distribution of X is bounded.

4.2. Piecewise polynomial regression

Piecewise polynomial estimators can be defined in different ways as they depend on a partition of the underlying space. Therefore we first have to describe the type of piecewise polynomials that we consider in the following.

Let $\hat{A} \in \mathbb{R}^{D\times d}$ contain column-wise an arbitrary orthonormal basis of $\mathrm{Im}(\hat{P})$. Denote by Δ_l the set of dyadic cubes in $\mathbb{R}^d$, *i.e.* the set of cubes with side length 2^{-l} and corners in the set $\{2^{-l}(v_1, \ldots, v_d) : v_j \in \mathbb{Z}\}$, and let $\Delta_l(R) \subseteq \Delta_l$ be the subset that has non-empty intersection with $\{\hat{A}^\top z : z \in B_R\}$, where $B_R = \{X \in \mathbb{R}^D : \|X\| \leq R\}$. Moreover, let $\mathcal{P}_k$ be the space of polynomials of degree k in $\mathbb{R}^d$ and 1_A be the characteristic function of a set A. The function space of piecewise polynomials we consider is defined by

$$\mathcal{F}(\hat{A}, l, k, R) := \left\{ f : f(x) = 1_{B_R}(x) \sum_{c\in\Delta_l(R)} 1_c\left(\hat{A}^\top x\right) p_c\left(\hat{A}^\top x\right), p_c \in \mathcal{P}_k \right\}.$$

To construct the estimator, we perform empirical risk minimization

$$\tilde{f} := \operatorname*{argmin}_{h\in\mathcal{F}(\hat{A},l,k,R)} \sum_{i=1}^{N} (h(X_i) - Y_i)^2,$$

and then set $\hat{f}(x) := T_{[-1,1]}(\tilde{f}(x))$, where $T_{[-1,1]}(u) := \mathrm{sign}(u)(|u| \wedge 1)$. Note that piecewise polynomial estimators are typically analyzed after thresholding to avoid technical difficulties with potentially unbounded predictions (see also [7, 19]).

The following theorem characterizes the influence of $\|\hat{P} - P\|$ on the generalization performance of the estimator.

Theorem 14. *Let g be (L, s)-smooth with $s = s_1 + s_2$, $s_1 \in \mathbb{N}_0$, $s_2 \in (0, 1]$. Choosing $l = \lceil \log_2(N)/(2s + d) \rceil$, $R^2 = D\,\|X\|^2_{\psi_2} \log(N)$, and $k = s_1$ we get*

$$MSE\left(\hat{f}, f\right) \leq C_1 \log^{1\vee\frac{d}{2}}(N) N^{-\frac{2s}{2s+d}} + C_2 \log(N)^{1\wedge s} \left\|\hat{P} - P\right\|^{2\wedge 2s}, \tag{23}$$

where the constants grow with σ_ζ, d, s, $L^* := Ld^{s_1/2}(1 - \|\hat{P} - P\|^2)^{-s/2}$, and C_1 depends linearly on $(D\,\|X\|^2_{\psi_2})^{d/2}$, and C_2 linearly on $(D\,\|X\|^2_{\psi_2})^{1\wedge s}$.

Remark 15 (Boundedness and log-factors)**.** For bounded X, the choice $R^2 \asymp \log(N)$ is not required and $\log^{1\vee\frac{d}{2}}(N)$ reduces to $\log(N)$. Moreover, $D\,\|X\|_{\psi_2}^2$ can be replaced by the squared radius of a ball containing the support of X, which removes the dependency on D entirely.

Remark 16 (Rate optimality)**.** Assuming $\|\hat{P}-P\| \in \mathcal{O}(N^{-1/2})$, we observe that the last term in (23) has order $N^{-s\wedge 1}$. Therefore, Theorem 14 ensures, up to log-factors, the optimal rate $N^{-2s/(2s+d)}$ for $d=1$ and $s \geq \frac{1}{2}$, or $d \geq 2$ and $s > 0$.

Proof sketch The first step is to apply the following well-known result.

Theorem 17 (Theorem 11.3 in [19])**.** *Let $\mathcal{F}$ be a vector space of functions $f : \mathbb{R}^D \to [-1,1]$. Assume $Y = f(X)+\zeta$, $\mathbb{E}[Y|X] = f(X)$ and $Var(\zeta|X=x) \leq \sigma_\zeta^2$. Denote by $\tilde{f}$ the empirical risk minimizer in $\mathcal{F}$ over N iid. copies of (X,Y), and let $\hat{f} = T_{[-1,1]}(\tilde{f})$. Then there exists a universal constant C such that*

$$MSE\left(\hat{f},f\right) \leq C(\sigma_\zeta^2 \vee 1)\frac{\log(N)+\dim(\mathcal{F})}{N} + C\inf_{h\in\mathcal{F}} MSE(h,f)\,. \tag{24}$$

The first term in (24) is the estimation error, which measures the deviation of the performance of the empirical risk minimizer to the best performing estimator in $\mathcal{F}$ when having access to the entire distribution. It decreases as more samples become available, but increases with the complexity of $\mathcal{F}$, here measured in terms of the dimensionality. It can be checked that $\mathcal{F}(\hat{A},l,k,R)$ is closed under addition and scalar multiplication and is thus a vector space. A basis can be constructed by combining the standard polynomial basis for each cell of the partition. Therefore $\dim(\mathcal{F}(\hat{A},l,k,R)) = |\Delta_l(R)|\binom{d+k}{k}$, where $|\Delta_l(R)|$ is the number of cells required to cover $\{\hat{A}^\top z : z \in B_R\}$. Lemma 25 in the Appendix proves $|\Delta_l(R)| \leq \lceil(2^{l+1}R)^d\rceil$ and therefore

$$\dim(\mathcal{F}(\hat{A},l,k,R)) \leq \binom{d+k}{k}\lceil(2^{l+1}R)^d\rceil. \tag{25}$$

The second term in (24) is the approximation error, which measures how well f can be approximated by any function $h \in \mathcal{F}(\hat{A},l,k,R)$. Neglecting for a moment the perturbation $\hat{P}-P$, it is known that a piecewise Taylor expansion of g can be used to approximate g with an accuracy that increases as the underlying partition is refined. The main difficulty in our case is to define a piecewise polynomial function $h \in \mathcal{F}(\hat{A},l,k,R)$ that approximates $g(A^\top x)$, despite the fact that h depends on coordinates $\hat{A}^\top x$ instead of $A^\top x$.

To define such a function, we first prove the existence of a function g^* that approximates g uniformly well, when being evaluated on $\hat{A}^\top x$. Precisely, Lemma 26 in the Appendix shows

$$\left|g^*(\hat{A}^\top x) - g(A^\top x)\right| \leq L^*\,\|x\|^{1\wedge s}\,\|\hat{P}-P\|^{1\wedge s},$$

for some (L^*, s)-smooth function g^*. Now, by approximating g^* through a piecewise Taylor expansion, we can construct a function $h \in \mathcal{F}(\hat{A}, l, k, R)$ which, using choices l, k and R as in Theorem 14, satisfies

$$\mathrm{MSE}\,(h, f) \leq C_1 N^{-\frac{2s}{2s+d}} + C_2 \log^{1\wedge s}(N) \left\| \hat{P} - P \right\|^{2\wedge 2s}, \tag{26}$$

for constants C_1 depending on L^*, d, s, and C_2 depending on L^* and linearly on $(D \|X\|_{\psi_2}^2)^{1\wedge s}$ (see Corollary 28). The proof of Theorem 14 concludes by combining Theorem 17, the dimensionality bound (25), and the approximation error bound (26) (see Appendix A.4).

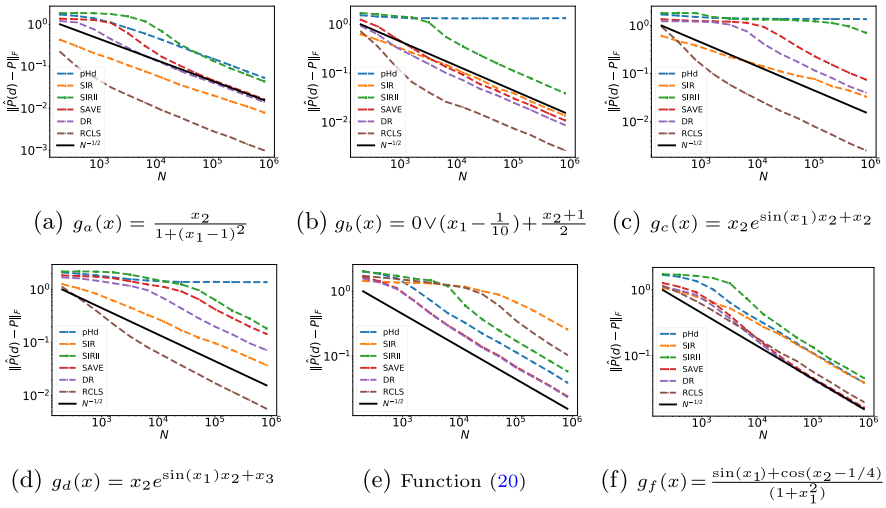
5. Numerical experiments

We now compare RCLS to the most prominent inverse regression based techniques SIR, SIRII, SAVE, pHd and DR that have been described extensively in Section 1.1. In the first part we consider synthetic problems and we directly assess the performance by evaluating $\|\hat{P}_J(d) - P\|_F$, since the true index space is known. In the second part, we consider real data sets from the UCI data set repository. Here, the true index space is unknown, and we instead compare recovered spaces $\mathrm{Im}(\hat{P})$ by measuring the predictive performance of kNN-regression, when trained on $\{(\hat{P}X_i, Y_i) : i \in [N]\}$. In both cases we construct the partition $\mathcal{R}_{J,\ell}$ using an equisized partition of the empirical range $[\min_i Y_i, \max Y_i]$ as described in Remark 1. The source code for all experiments is readily available at `https://github.com/soply/sdr_toolbox` and `https://github.com/soply/mim_experiments`.

5.1. Synthetic data sets

We sample $X \sim \mathrm{Uni}(\{X : \|X\| \leq 1\})$ in $\mathbb{R}^{20}$, and generate the response by $Y = g(A^\top X) + \zeta$ for several functions $g \in \{g_a, \ldots, g_f\}$, each one defined below the respective plot in Figure 3, and $\zeta \sim \mathcal{N}(0, 0.01^2 \mathrm{Var}(g(A^\top X)))$. The index space is $A = [e_1 | \ldots | e_d] \in \mathbb{R}^{D\times d}$ where e_i is the i-th standard basis vector. The hyperparameter J is chosen optimally for SIR, SIRII, SAVE, DR and RCLS to minimize the projection error within $J \in [100]$. No parameter is required for pHd.

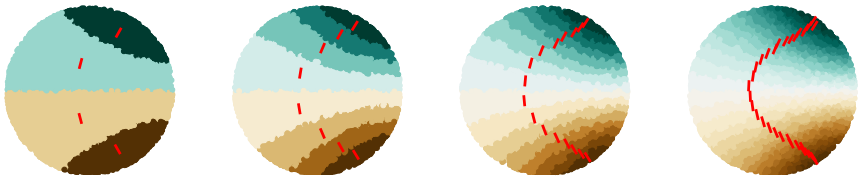
We report projection errors averaged over 100 repetitions of the same experiment in Figures 3a - 3f. First, notice that most estimators (except pHd and SIR in some cases) achieve the expected $N^{-1/2}$ rate on all problems. pHd fails to detect linear trends and therefore fails to detect the index space in some cases. RCLS achieves the best performance in Figures 3a - 3d, performs poorly on 3e, and is tied with SAVE and DR on Figure 3f. The poor performance in 3e is related to the fact that the ranges $\mathrm{Im}(h_1), \mathrm{Im}(h_2), \mathrm{Im}(h_3)$ are almost completely overlapping (see (20)), which means that most of the conditional distributions $(X_{J,\ell}, Y_{J,\ell})$, $\ell \in [J]$, generate the same direction, and the remaining two directions are only visible on a small fraction of the distribution (X, Y).



(a) $g_a(x) = \frac{x_2}{1+(x_1-1)^2}$ (b) $g_b(x) = 0 \vee (x_1 - \frac{1}{10}) + \frac{x_2+1}{2}$ (c) $g_c(x) = x_2 e^{\sin(x_1)x_2+x_2}$

(d) $g_d(x) = x_2 e^{\sin(x_1)x_2+x_3}$ (e) Function (20) (f) $g_f(x) = \frac{\sin(x_1)+\cos(x_2-1/4)}{(1+x_1^2)}$

FIG 3. *Projection error $\|\hat{P}_J(d) - P\|_F$ versus sample size N for different methods and functions. For all estimators except pHd and SIR, the expected $N^{-1/2}$ rate is observed on all problems. In some cases, RCLS stands out by temporarily attaining faster decay and obtaining significantly smaller error.*



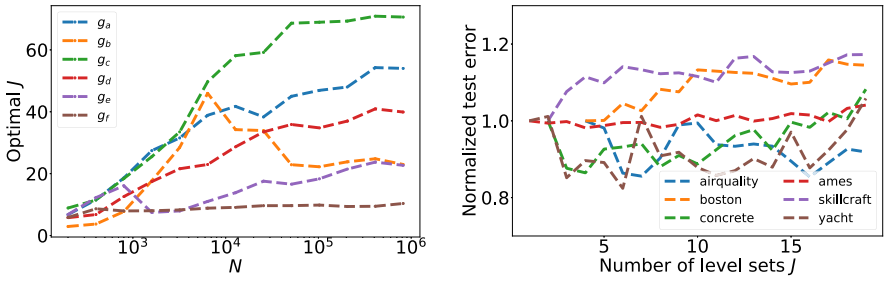
(a) $J = 4$ (b) $J = 8$ (c) $J = 16$ (d) $J = 32$

FIG 4. *Level set partition of $\{X : \|X\| \leq 1\} \subset \mathbb{R}^2$ induced by equisized intervals $\mathcal{R}_{J,\ell}$ for link function $g_a(x) = \frac{x_2}{1+(x_1-1)^2}$. Different colors represent different level sets and red bars are local vectors $b_{J,\ell}$.*

In Figures 3a - 3d, where RCLS improves upon competitors, we observe temporary error decays beyond the rate $N^{-1/2}$ from Corollary 8. A possible explanation for this phenomenon is given in Remark 9, which shows that convergence rates up to N^{-1} are possible in an idealized scenario and when choosing the number of level sets J as a monotone increasing function of N. Judging by Figure 5a, the optimal J indeed exhibits a fairly monotone increase with the number of samples N for link functions g_a, g_b, g_c, g_d, at least within a certain range of N's. As pointed out in Remark 9, a second requirement for faster error decay seems to be that local covariances behave roughly like

$$\Sigma_{J,\ell} \approx J^{-2} P_{J,\ell} + Q_{J,\ell}. \tag{27}$$

We believe this can be a reasonable approximation if X follows a generic distributions, *e.g.* $X \sim \text{Uni}(\{X : \|X\| \leq 1\})$ or $X \sim \mathcal{N}(0, \mathsf{Id}_D)$, and if the function



(a) (b)

FIG 5. *(a) Optimal choices for parameter J for the RCLS method in all synthethic examples evaluated in Figure 3. The range $J \in [100]$ was used in experiments. (b) Sensitivity analysis of RCLS with respect to J on all real data sets. We fix the subspace dimension d and the number of neighbors k to the nearest integer to cross-validated values in Table 2 and plot the error, averaged over* 100 *repititions, as a function of parameter J. The error is normalized so that it equals* 1 *for the choice $J = d$.*

follows locally (in response domain) monotone single-index model structure. In such cases, the resulting partition qualitatively should look like the one in Figure 4, which is generated by $g_a(x) = \frac{x_2}{1+(x_1-1)^2}$ and $J \in \{4, 8, 16, 32\}$. Lastly, we also add that faster convergence rates with an estimator based on response-conditional least squares vectors have also been observed and analyzed in [27, 28] for the monotone single-index model, and a nonlinear generalizations thereof.

5.2. Real data sets

To compare RCLS with inverse regression based competitors on real data sets, we first compute an index space and then compare the predictive performance when training a kNN-regressor on projected samples. More precisely, we conduct the following steps.

1. Split the data set $\{(X_i, Y_i) : i \in [N]\}$ into training and test set $\mathcal{X}_{\text{Train}}$, $\mathcal{Y}_{\text{Train}}$, and $\mathcal{X}_{\text{Test}}, \mathcal{Y}_{\text{Test}}$
2. Use pHd, SIR, SIRII, SAVE, DR, RCLS on the training set to compute an index space $\hat{A}$
3. Train a kNN-regressor using $\{(\hat{A}^\top X_i, Y_i) : X_i \in \mathcal{X}_{\text{Train}}\}$
4. Crossvalidate over hyperparameters d (index space dimension), k (kNN parameter), and J (number of level sets) using a hold-out validation set of the training data
5. Compute the root mean squared error (RMSE) of the kNN-regressor on the test set

The test set contains 15% of the data, while cross-validation is performed using a 10-fold splitting strategy. Each experiment is repeated 20 times and we report the mean and standard deviation.

TABLE 2

RMSE, standard deviation, and cross-validated hyperparameters, over 20 *repetitions for several estimators and UCI repository data sets. Values for d, k, J are averages over different runs of each experiment. First* 5 *rows describe the data sets and their characteristics, and the remaining rows contain the results. For a simplified presentation, we divide the mean and STD of RMSE, and the mean and STD of the data (5th row) by the value in row* Factor.

Characteristics	Airquality	Ames	Boston	Concrete	Skillcraft	Yacht
log-TF	No	Yes	Yes	No	Yes	Yes
$D,\ N$	11, 7393	7, 1197	12, 506	8, 1030	16, 3338	6, 307
Factor	10^{-1}	10^{5}	10^{1}	10^{1}	10^{2}	10^{1}
$\bar{Y} \pm \mathrm{STD}(Y)$	9.95 ± 4.03	1.74 ± 0.67	1.27 ± 0.71	3.58 ± 1.67	1.15 ± 0.48	1.05 ± 1.51
Baselines						
LinReg	1.22 ± 0.03	$\mathbf{0.23 \pm 0.02}$	0.50 ± 0.11	1.06 ± 0.06	0.14 ± 0.03	0.22 ± 0.07
kNN	1.03 ± 0.02	0.26 ± 0.03	0.41 ± 0.06	0.89 ± 0.08	0.17 ± 0.01	0.76 ± 0.11
k	25.0	9.8	6.8	5.5	9.8	1.1
SDR + KNN						
pHd	1.01 ± 0.04	0.29 ± 0.04	0.43 ± 0.06	0.89 ± 0.04	0.26 ± 0.01	0.70 ± 0.18
d	10.2	6.5	7.55	6.65	7.0	5.4
k	10.0	7.7	5.3	4.0	16.4	1.7
SIR	0.59 ± 0.02	0.24 ± 0.04	0.42 ± 0.07	0.87 ± 0.06	0.08 ± 0.01	0.18 ± 0.07
d	6.15	1.4	4.8	6.1	1.85	1.15
k	10.0	11.4	6.95	3.7	11.4	8.4
J	10.3	10.8	8.6	9.8	7.65	8.6
SIRII	0.87 ± 0.02	0.27 ± 0.04	0.44 ± 0.07	0.89 ± 0.06	0.29 ± 0.03	0.53 ± 0.21
d	11.0	7.0	6.65	8.0	2.75	2.8
k	10.0	8.45	5.3	4.0	16.75	6.7
J	12.0	9.8	8.2	10.4	3.75	4.6
SAVE	0.58 ± 0.01	0.25 ± 0.04	0.44 ± 0.05	0.79 ± 0.09	0.09 ± 0.01	0.18 ± 0.04
d	6.0	2.9	7.55	3.85	4.65	1.7
k	10.0	10.7	6.35	3.2	9.5	7.35
J	13.4	12.0	9.9	8.3	8.75	8.7
DR	0.58 ± 0.02	0.24 ± 0.03	$\mathbf{0.40 \pm 0.07}$	0.75 ± 0.08	0.08 ± 0.01	0.18 ± 0.06
d	5.85	2.15	5.45	3.35	2.6	1.75
k	10.0	11.3	6.8	3.7	14.9	6.55
J	12.6	12.0	10.2	8.5	9.3	8.75
RCLS	$\mathbf{0.51 \pm 0.03}$	0.25 ± 0.03	0.41 ± 0.06	$\mathbf{0.72 \pm 0.06}$	$\mathbf{0.07 \pm 0.01}$	$\mathbf{0.17 \pm 0.06}$
d	5.4	1.9	5.1	3.25	2.95	1.75
k	10.0	13.55	11.0	4.3	9.15	5.6
J	11.9	6.4	5.7	5.7	3.2	7.3

We consider the UCI data sets `Airquality`, `Ames`-housing, `Boston`-housing, `Concrete`, `Skillcraft` and `Yacht`. We standardize the components of X to $[-1, 1]$ and potentially perform a log transformation of Y if the marginal has sparsely populated tails. This is indicated by the log-*TF* row in Table 2. For some data sets, we also exclude features with missing values, or, in the case of `Ames`, we exclude some irrelevant and categorical features to reduce the complexity of the data set. Preprocessed data sets can be found at https://github.com/soply/db_hand.

The RMSE and cross-validated hyperparameters are presented in Table 2. To have robust baselines for comparison, we also compute the RMSE of standard linear regression and kNN regression. We first see that applying a dimension reduction technique improves the performance of linear regression and ordinary kNN significantly on data sets *Airquality*, *Concrete*, *Skillcraft* and *Yacht*. Furthermore, on these data sets, RCLS convinces by achieving the best results among all competitors. Runner-up is DR, where SIR and SAVE share third and fourth place. The results of pHd and SIRII are not convincing on most data sets.

The sensitivity of RCLS with respect to parameter J is illustrated in Figure 5b. Respective test errors between best and worst fit vary by around up to 20%, indicating that the choice of the partition, as for most inverse regression techniques, plays an important role in RCLS. Fortunately, the efficiency of the method allows for fast cross-validation and trying different partitioning strategies as mentioned in Remark 1.

The study confirms that RCLS is a viable alternative to other prominent inverse regression methods. Some of the data sets were chosen because one-dimensional maps $e_i^\top X \mapsto Y$, where e_i is the i-th standard basis vector, show a certain degree of monotonicity. We believe this could promote local monotone single-index model structure, respectively, that partitioning affects local covariances according to the model (27), and thus may be beneficial for the performance of RCLS.

Appendix

A.1. Probabilistic results

This section contains some probabilistic auxiliary results used in the paper.

Lemma 18. *If $Z \in \mathbb{R}^D$ is sub-Gaussian and E an event with $\mathbb{P}(E) > 0$, then $Z|E$ is sub-Gaussian with $\|Z|E\|_{\psi_2} \leq \|Z\|_{\psi_2} \mathbb{P}(E)^{-1}$.*

Proof. Assume without loss of generality $Z \in \mathbb{R}$. The result for the vector then follows by the definition. We use the characterization of sub-Gaussianity by the moment bound in [54, Proposition 2.5.2, b)]. So let $p \geq 1$. By the law of total expectation it follows that $\mathbb{E}[|Z|^p] = \mathbb{E}[|Z|^p\,|E]\mathbb{P}(E) + \mathbb{E}[|Z|^p\,|E^C]\mathbb{P}(E^C) \geq \mathbb{E}[|Z|^p\,|E]\mathbb{P}(E)$. Dividing $\mathbb{P}(E)$ and using monotonicity of the p-th root

$$\left(\mathbb{E}[|Z|^p\,|E]\right)^{1/p} \leq \frac{\left(\mathbb{E}[|Z|^p]\right)^{1/p}}{\mathbb{P}(E)^{1/p}} \leq \frac{\left(\mathbb{E}[|Z|^p]\right)^{1/p}}{\mathbb{P}(E)} \leq C\frac{\|Z\|_{\psi_2}\sqrt{p}}{\mathbb{P}(E)},$$

where C is some universal constant, the second inequality follows from $\mathbb{P}(E) \leq 1$, and the third from the sub-Gaussianity of Z. □

Lemma 19. *If $X \in \mathbb{R}^D$ is sub-Gaussian, so is $\|X\|$, with $\|\|X\|\|_{\psi_2} \leq \sqrt{D}\|X\|_{\psi_2}$.*

Proof. Using Hölder's inequality and the sub-Gaussianity of X we compute

$$\mathbb{E}\left[\exp\left(\frac{\|X\|^2}{D\|X\|_{\psi_2}^2}\right)\right] = \mathbb{E}\left[\prod_{i=1}^{D}\exp\left(\frac{|e_i^T X|^2}{D\|X\|_{\psi_2}^2}\right)\right]$$
$$\leq \left(\prod_{i=1}^{D}\mathbb{E}\left[\exp\left(\frac{|e_i^T X|^2}{\|X\|_{\psi_2}^2}\right)\right]\right)^{1/D} \leq 2. \qquad \square$$

Lemma 20. *Fix $u > 0$, $\varepsilon > 0$. Let $Y \in \mathbb{R}$ be a random variable, $\mathcal{R}$ an interval, and $\hat{\mathbb{P}}(Y \in \mathcal{R}) := |\{Y_i \in \mathcal{R}\}|\, N^{-1}$ the empirical estimate of $\mathbb{P}(Y \in \mathcal{R})$ based on N iid. samples. Then, with probability at least $1 - \exp(-u)$ we have*

$$\left|\hat{\mathbb{P}}(Y \in \mathcal{R}) - \mathbb{P}(Y \in \mathcal{R})\right| \leq \sqrt{\mathbb{P}(Y \in \mathcal{R})}\left(\varepsilon \wedge \frac{1}{2}\sqrt{\mathbb{P}(Y \in \mathcal{R})}\right) \tag{28}$$

provided $N > Cu(\mathbb{P}(Y \in \mathcal{R})^{-1} \vee \varepsilon^{-2})$.

Proof. Let $\hat{p} := \hat{\mathbb{P}}(Y \in \mathcal{R})$ and $p = \mathbb{P}(Y \in \mathcal{R})$ for short and define $\delta = (\varepsilon/\sqrt{p} \wedge 1/2) \in (0, 1/2)$. We can use a Chernoff bound from [47] for the random variable $N\hat{p}$ with expectation Np to get

$$\mathbb{P}\left(\hat{p} - p \leq -\delta p\right) = \mathbb{P}\left(N\hat{p} - Np \leq -\delta Np\right) \leq \exp\left(-\frac{\delta^2 Np}{2}\right),$$
$$\text{and } \mathbb{P}\left(\hat{p} - p \geq \delta p\right) = \mathbb{P}\left(N\hat{p} - Np \geq \delta Np\right) \leq \exp\left(-\frac{\delta^2 Np}{2+\delta}\right) \leq \exp\left(-\frac{\delta^2 Np}{5/2}\right).$$

Taking the union bound over both events we therefore have

$$\mathbb{P}\left(|\hat{p} - p| \leq \delta p\right) \geq 1 - 2\exp\left(-\frac{\delta^2 Np}{5/2}\right) = 1 - 2\exp\left(-\frac{\min\{\varepsilon^2, p\}N}{10}\right).$$

The result follows by the condition on N assuming large enough $C > 0$. $\square$

A.2. Differences of projections

We gather two auxiliary results to rewrite the norm of differences of projections.

Lemma 21. *Let A and B be subspaces with $\dim(A) = \dim(B)$, and let P_A and P_B the corresponding orthogonal projections. For $P_{A^\perp} = \mathsf{Id} - P_A$ we get $\|P_A - P_B\| = \|P_{A^\perp} P_B\|$.*

Proof. Assume $\|(\mathsf{Id} - P_A)\, P_B\| = \|P_{A^\perp} P_B\| < 1$ first. Then the first case of Theorem 6.34 in Chapter 1 in [26] applies. Note that the second case can be ruled out since P_A can not map $\mathrm{Range}(P_B)$ one-to-one onto a proper subspace of $V \subset \mathrm{Range}(P_A)$ because $\dim(V) < \dim(\mathrm{Range}(P_A)) = \dim(\mathrm{Range}(P_B))$ according to the assumption. Thus, in the first case it follows that

$$\|(\mathsf{Id} - P_A)\, P_B\| = \|(\mathsf{Id} - P_B)\, P_A\| = \|P_A - P_B\|.$$

Now let $\|P_{A^\perp} P_B\| = 1$. Then there exists $v \in \mathbb{S}^D$ such that $\|P_{A^\perp} P_B v\| = \|v\|$. Since

$$\|v\| \geq \|P_B v\| \geq \|P_{A^\perp} P_B v\| = \|v\|,$$

it follows that $\|P_B v\| = \|v\|$, and thus $P_B v = v$ because P_B is a projection. With the same argument we deduce also $P_{A^\perp} v = v$, and then

$$(P_A - P_B)\, v = P_A v - P_B v = 0 - v = v$$

implies $\|P_A - P_B\| = 1 = \|P_{A^\perp} P_B\|$. □

Lemma 22. *Let A and B be subspaces with $m = \dim(A) = \dim(B)$, and let P_A and P_B the corresponding orthogonal projections. For $P_{A^\perp} = \mathsf{Id} - P_A$ we get $\|P_A - P_B\|_F = \sqrt{2}\,\|P_{A^\perp} P_B\|_F$.*

Proof. With slight abuse of notation we denote $A, B \in \mathbb{R}^{D \times m}$ two orthonormal bases of A respectively B such that $P_A = AA^\top$, $P_B = BB^\top$. Now, denote $A^\top B = U(cos(\theta))V^\top$ where $\cos(\theta) \in \mathbb{R}^{m \times m}$ is the diagonal matrix containing the principal angles θ_i [20]. From [20] we obtain the identity $1/2\,\|P_A - P_B\|_F^2 = m - \sum_{i=1}^m \cos^2(\theta_i)$. Doing some further manipulations we get

$$\begin{aligned}
\frac{1}{2}\,\|P_A - P_B\|_F^2 &= m - \sum_{i=1}^m \cos^2(\theta_i) = m - \left\|A^\top B\right\|_F^2 \\
&= m - \mathrm{Trace}(B^\top A A^\top B) = m - \mathrm{Trace}(AA^\top BB^\top) \\
&= m - \mathrm{Trace}(P_A P_B) = m - \mathrm{Trace}((\mathsf{Id} - P_{A^\perp})P_B) \\
&= m - \mathrm{Trace}(P_B) + \mathrm{Trace}(P_{A^\perp} P_B).
\end{aligned}$$

The result follows by $\mathrm{Trace}(P_B) = m$ and $\mathrm{Trace}(P_{A^\perp} P_B) = \|P_{A^\perp} P_B\|_F^2$. □

A.3. Proof of Theorem 11

Proof of Theorem 11. Denote $S_X = \{X_i : i \in [N]\}$ and $\tilde{f}_k(x) = \mathbb{E}[\hat{f}_k(x)|S_X] = \sum_{i=1}^k f(X_{i(x)})$ for fixed x. We first decompose (randomness is in the ζ_i's)

$$\mathbb{E}\left[\left(\hat{f}_k(x) - f(x)\right)^2 \middle| S_X\right] = \mathbb{E}\left[\left(\hat{f}_k(x) - \tilde{f}_k(x)\right)^2 \middle| S_X\right] + \left(\tilde{f}_k(x) - f(x)\right)^2,$$

and then use the towering property of conditional expectations to obtain

$$\begin{aligned}
\mathbb{E}\left(\hat{f}_k(X) - f(X)\right)^2 &= \mathbb{E}\mathbb{E}\left[\left(\hat{f}_k(X) - f(X)\right)^2 \middle| \mathcal{S}_X\right] \\
&= \mathbb{E}\mathbb{E}\left[\left(\hat{f}_k(X) - \tilde{f}_k(X)\right)^2 \middle| \mathcal{S}_X\right] + \mathbb{E}\left(\tilde{f}_k(X) - f(X)\right)^2 \\
&= \mathbb{E}\left(\hat{f}_k(X) - \tilde{f}_k(X)\right)^2 + \mathbb{E}\left(\tilde{f}_k(X) - f(X)\right)^2
\end{aligned}$$

Since $\mathbb{E}\zeta_i = 0$, $\zeta_i \perp\!\!\!\perp \zeta_j$ and $\mathrm{Var}(\zeta_i | X = x) \leq \sigma_\zeta^2$, the first term satisfies the bound

$$\mathbb{E}\left(\hat{f}_k(X) - \tilde{f}_k(X)\right)^2 = \mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^{k} \zeta_{i(X)}\right) \leq \frac{\sigma_\zeta^2}{k} = C_k^{-1}\sigma_\zeta^2 N^{-\frac{2s}{2s+d}}.$$

For $\mathbb{E}\left(\tilde{f}_k(X) - f(X)\right)^2$, we recall that $f(X) = g(A^\top X)$ for some (L,s)-smooth g, which implies

$$\begin{aligned}\mathbb{E}\left(\tilde{f}_k(X) - f(X)\right)^2 &= \mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k} g(A^\top X_{i(X)}) - g(A^\top X)\right)^2 \\ &\leq 4L^2\mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k} \min\left\{\left\|A^\top\left(X_{i(X)} - X\right)\right\|^s, 1\right\}\right)^2,\end{aligned}$$

where the 4 can be injected since $|f(X)| \leq 1$ almost surely. To bound this term further we have to replace $\{X_{i(X)} : i \in [k]\}$ (the k closest samples wrt to $\hat{d}(\cdot, X) := \|\hat{P}(\cdot - X)\|$) by the k closest samples based on $d(\cdot, X) := \|P(\cdot - X)\|$. So let $\tilde{X}_{i(X)}$ denote the i-th closest sample to X based on d, and let further $\delta := \left\|\hat{P} - P\right\|$. Since

$$\left|d(X, X') - \hat{d}(X, X')\right| \leq \left\|(P - \hat{P})(X - X')\right\| \leq \delta\|X - X'\|,$$

and $(a+b)^s \leq a^s + b^s$ for $s \leq 1$, we can bound

$$\begin{aligned}\sum_{i=1}^{k} \min\{d(X_{i(X)}, X)^s, 1\} &\leq \sum_{i=1}^{k} \min\{\hat{d}(X_{i(X)}, X)^s + \delta^s \max_{i\in[N]}\|X_i - X\|^s, 1\} \\ &\leq \sum_{i=1}^{k} \min\{\hat{d}(\tilde{X}_{i(X)}, X)^s + \delta^s \max_{i\in[N]}\|X_i - X\|^s, 1\} \\ &\leq \sum_{i=1}^{k} \min\left\{d(\tilde{X}_{i(X)}, X)^s + 2\delta^s \max_{i\in[N]}\|X_i - X\|^s, 1\right\},\end{aligned}$$

where in the second inequality we used that $X_{i(X)}$ minimizes the distance to X measured in $\hat{d}$, and can therefore be replaced by $\tilde{X}_{i(X)}$. Denoting $\Delta_{X,N} := 2\delta^s \max_{i\in[N]}\|X_i - X\|^s$ and using $(\sum_{i=1}^k b_i)^2 \leq k\sum_{i=1}^k b_i^2$ for arbitrary b_i's, we get

$$\begin{aligned}\mathbb{E}\left(\hat{f}_k(X) - f(X)\right)^2 &\leq 4L^2\mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k} \min\{d(\tilde{X}_{i(X)}, X)^s + \Delta_{X,N}, 1\}\right)^2 \\ &\leq \frac{4L^2}{k}\sum_{i=1}^{k} \mathbb{E}\min\{(d(\tilde{X}_{i(X)}, X)^s + \Delta_{X,N})^2, 1\} \qquad (29) \\ &\leq \frac{8L^2}{k}\sum_{i=1}^{k} \mathbb{E}\min\{d(\tilde{X}_{i(X)}, X)^{2s}, 1\} + \frac{8L^2}{k}\sum_{i=1}^{k}\mathbb{E}\Delta_{X,N}^2.\end{aligned}$$

For the first term, we proceed as in [19, 29] by randomly splitting the data set $\{X_i : i \in [N]\}$ into $k+1$ sets, where the first k sets contain $\lfloor N/k \rfloor$ samples. Then we let $X^*_{i(X)}$ denote the nearest neighbor to X (measured in d) within the i-th set. Since $\{\tilde{X}_{i(X)} : i \in [k]\}$ are by definition the closest k samples (measured in d), we can bound

$$\begin{aligned}\frac{8L^2}{k}\sum_{i=1}^{k}\mathbb{E}\min\{d(\tilde{X}_{i(X)},X)^{2s},1\} &\le \frac{8L^2}{k}\sum_{i=1}^{k}\mathbb{E}\min\{d(X^*_{i(X)},X)^{2s},1\}\\ &= 8L^2\mathbb{E}\min\left\{\left\|P\left(X^*_{1(X)}-X\right)\right\|^{2s},1\right\},\end{aligned}$$

where the last equality uses that the distribution of $X^*_{i(X)} - X$ is independent of the set index i. Since $\|P(\cdot)\| = \left\|A^\top(\cdot)\right\|$, $d > 2s$ by assumption, and

$$\mathbb{E}\left\|A^\top X\right\|^\beta = \mathbb{E}\|X\|^\beta \lesssim \|X\|^\beta_{\psi_2}\,\beta^{\beta/2}$$

for any $\beta \ge 1$ by the sub-Gaussianity of X (see [54, Proposition 2.5.2]), Lemma 1 in [29] implies the existence of a constant $C_1 = C_1(d, s, \|X\|_{\psi_2})$ satisfying

$$\begin{aligned}\mathbb{E}\min\left\{\left\|P\left(X^*_{1(X)}-X\right)\right\|^{2s},1\right\} &= \mathbb{E}\min\left\{\left\|A^\top\left(X^*_{1(X)}-X\right)\right\|^{2s},1\right\}\\ &\le C_1\left(\frac{k}{N}\right)^{\frac{2s}{d}} = C_1 C_k^{\frac{2s}{d}} N^{-\frac{2s}{2s+d}}.\end{aligned}$$

It remains to bound the last term in (29). Denote for short $\sigma_X = \|X\|_{\psi_2}$. We first compute that

$$\mathbb{E}\Delta^2_{X,N} = \int_0^\infty \mathbb{P}\left(\Delta^2_{X,N} > u\right)du = \int_0^\infty \mathbb{P}\left(\max_{i=1,\dots,N}\|X_i - X\|^{2s} > \frac{u}{4\delta^{2s}}\right)du.$$

We can control this probability by using the sub-Gaussianity of X. More precisely, since X is sub-Gaussian, $X_i - X$ is sub-Gaussian (norm changing only by a universal constant), and Lemma 19 implies that $\|\,\|X_i - X\|\,\|_{\psi_2} \le C\sqrt{D}\sigma_X$. Taking the square, and using [54, Lemma 2.7.6], we obtain

$$\|\,\|X_i - X\|^2\,\|_{\psi_1} \le \|\,\|X_i - X\|\,\|^2_{\psi_2} \le CD\sigma_X^2.$$

To bound the integral, we first split $[0,\infty]$ into intervals $[0, \nu D\sigma_X^2\log(N)\delta^{2s}]$ and $[\nu D\sigma_X^2\log(N)\delta^{2s}, \infty]$ for some $\nu > 4\max\{\frac{1}{D\sigma_X^2\log(N)}, C\}$, which yields

$$\mathbb{E}\Delta^2_{X,N} \le \int_{\nu D\sigma_X^2\log(N)\delta^{2s}}^\infty \mathbb{P}\left(\max_{i\in[N]}\|X_i - X\|^{2s} > \frac{u}{4\delta^{2s}}\right)du + \nu D\sigma_X^2\log(N)\delta^{2s}.$$

For the first term we realize that $u > \nu D\sigma_X^2\log(N)\delta^{2s} > 4\delta^{2s}$ implies

$$\mathbb{P}\left(\max_{i=1,\dots,N}\|X_i - X\|^{2s} > \frac{u}{4\delta^{2s}}\right) \le \mathbb{P}\left(\max_{i=1,\dots,N}\|X_i - X\|^{2} > \frac{u}{4\delta^{2s}}\right).$$

Then the sub-Exponentiality of $\|X - X_i\|^2$ and a union bound argument over $i \in [N]$ give

$$\begin{aligned}
\mathbb{E}\Delta_{X,N}^2 &\leq \int_{\nu D\sigma_X^2 \log(N)\delta^{2s}}^{\infty} \mathbb{P}\left(\max_{i\in[N]} \|X_i - X\|^2 > \frac{u}{4\delta^{2s}}\right) du + \nu D\sigma_X^2 \log(N)\delta^{2s} \\
&\leq 2N \int_{\nu D\sigma_X^2 \log(N)\delta^{2s}}^{\infty} \exp\left(-\frac{u}{CD\sigma_X^2\delta^{2s}}\right) du + \nu 2D\sigma_X^2 \log(N)\delta^{2s} \\
&\leq 2CD\sigma_X^2\delta^{2s} N \exp\left(-\frac{\nu \log(N)}{C}\right) + \nu D\sigma_X^2 \log(N)\delta^{2s} \\
&\leq 2CD\sigma_X^2\delta^{2s} N \exp\left(-\log(N)\right) + \nu D\sigma_X^2 \log(N)\delta^{2s} \\
&\leq 2(\nu \vee C) D\sigma_X^2 \log(N)\delta^{2s}. \qquad \square
\end{aligned}$$

A.4. Proof of Theorem 14

Interlude: smoothness of linear concatenations In this section we establish smoothness properties of linear concatenations with explicit bounds for corresponding Lipschitz constants.

Lemma 23. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$, $W \in \mathbb{R}^{d\times d}$ and $\psi(z) = \phi(Wz)$. Let $s \in \mathbb{N}$ and $\alpha \in \mathbb{N}^d$ be a multi-index with $\sum_{i=1}^d \alpha_i = k \leq s$. If $\phi \in \mathcal{C}^s(\mathbb{R}^d)$, i.e. all partial derivatives $\partial^\alpha \phi$ exist and are continuous, then also all $\partial^\alpha \psi$ exist and are continuous. Moreover, if $i : [k] \to [d]$ is an arbitrary derivative ordering satisfying $\alpha = \sum_{w=1}^k e_{i(w)}$, we can express for any $k \in [s]$*

$$\partial^\alpha \psi(z) = \sum_{i_1=1}^d \cdots \sum_{i_k=1}^d \left(\prod_{w=1}^k W_{i_w, i(w)}\right) \partial_{i_1} \cdots \partial_{i_k}(\phi)(Wz).$$

Example 1. Let $\alpha = e_i + e_j$ and $i(1) = i$, $i(2) = j$. Then the formula yields the derivative

$$\partial^\alpha \psi(z) = \sum_{i_1=1}^d \sum_{i_2=1}^d W_{i_1,i} W_{i_2,j} \partial_{i_1}\partial_{i_2}(\phi)(Wz).$$

Proof. ψ is a concatenation of a $\mathcal{C}^s$ function with a linear transformation and is therefore as smooth as ϕ. For the formula, we use induction over k. Let α be a multi-index with $\sum_{i=1}^d \alpha_i = 1$, i.e. α is equal to a standard basis vector e_i for some $i \in [d]$. Since $\nabla\psi(z) = W^T \nabla\phi(Wz)$ we have

$$\partial_i \psi(z) = \langle W_i, \nabla\phi(Wz)\rangle = \sum_{i_1=1}^d W_{i_1,i} \partial_{i_1}(\phi)(Wz).$$

For the induction step $k-1 \to k$, we let α be a multi-index with $\sum_{i=1}^d \alpha_i = k$ and we calculate $\partial^\alpha \psi(z) = \partial_{i(k)} \partial^{\alpha - e_{i(k)}} \psi(z)$. Since $\alpha - e_{i(k)}$ is a multi-index

whose entries sum to $k-1$, by induction hypothesis we have

$$\begin{aligned}
\partial^\alpha \psi(z) &= \partial_{i(k)} \left(\sum_{i_1=1}^{d} \cdots \sum_{i_{k-1}=1}^{d} \left(\prod_{w=1}^{k-1} W_{i_w,i(w)} \right) \partial_{i_1} \cdots \partial_{i_{k-1}}(\phi)(Wz) \right) \\
&= \sum_{i_1=1}^{d} \cdots \sum_{i_{k-1}=1}^{d} \left(\prod_{w=1}^{k-1} W_{i_w,i(w)} \right) \partial_{i(k)} \left(\partial_{i_1} \cdots \partial_{i_{k-1}}(\phi)(Wz) \right) \\
&= \sum_{i_1=1}^{d} \cdots \sum_{i_{k-1}=1}^{d} \left(\prod_{w=1}^{k-1} W_{i_w,i(w)} \right) \sum_{i_k=1}^{d} W_{i_k,i(k)} \partial_{i_1} \cdots \partial_{i_k}(\phi)(Wz) \\
&= \sum_{i_1=1}^{d} \cdots \sum_{i_{k-1}=1}^{d} \sum_{i_k=1}^{d} \left(\prod_{w=1}^{k-1} W_{i_w,i(w)} \right) W_{i_k,i(k)} \partial_{i_1} \cdots \partial_{i_k}(\phi)(Wz),
\end{aligned}$$

where we used Schwartz Lemma in the second to last equality The result follows by extending the product. □

Lemma 24. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$, $s_1 \in \mathbb{N}_0$, $0 < s_2 \leq 1$ and $s = s_1 + s_2$. Assume ϕ is (L,s)-smooth, $W \in \mathbb{R}^{d\times d}$, and define $\psi(z) = \phi(Wz)$ for some $W \in \mathbb{R}^{d\times d}$. Then ψ is $(Ld^{\frac{s_1}{2}} \|W\|^s, s)$-smooth.*

Proof. Since W is a linear transformation, ψ has as many continuous partial derivatives as ϕ. Now consider $\alpha \in \mathbb{N}_0^d$ with $\sum_{i=1}^d \alpha_i = s_1$, and let $i : [s_1] \to [d]$ be an arbitrary derivative ordering satisfying $\sum_{w=1}^{s_1} e_{i(w)} = \alpha$. By using Lemma 23, we get

$$\begin{aligned}
&|\partial^\alpha \psi(z) - \partial^\alpha \psi(z')| \\
&= \left| \sum_{i_1=1}^{d} \cdots \sum_{i_{s_1}=1}^{d} \left(\prod_{w=1}^{s_1} W_{i_w,i(w)} \right) \left(\partial_{i_1} \cdots \partial_{i_{s_1}}(\phi)(Wz) - \partial_{i_1} \cdots \partial_{i_{s_1}}(\phi)(Wz') \right) \right| \\
&\leq \max_{\alpha : \sum_{i=1}^d \alpha_i = s_1} |\partial^\alpha \phi(Wz) - \partial^\alpha \phi(Wz')| \left(\sum_{i_1=1}^{d} \cdots \sum_{i_{s_1}=1}^{d} \prod_{w=1}^{s_1} \left| W_{i_w,i(w)} \right| \right).
\end{aligned}$$

Furthermore denote $\|W\|_1 = \max_i \sum_j |W_{i,j}| \leq \sqrt{d}\,\|W\|$. Then we can rewrite

$$\begin{aligned}
\sum_{i_1=1}^{d} \cdots \sum_{i_{s_1}=1}^{d} \prod_{w=1}^{s_1} \left| W_{i_w,i(w)} \right| &= \sum_{i_1=1}^{d} \left| W_{i_1,i(1)} \right| \cdots \sum_{i_{s_1}=1}^{d} \left| W_{i_{s_1},i(s_1)} \right| \\
&\leq \|W\|_1^{s_1} \leq d^{\frac{s_1}{2}} \|W\|^{s_1}.
\end{aligned}$$

Combining this with the previous calculation, and the fact that ϕ is (L,s)-smooth, we get

$$|\partial^\alpha \psi(z) - \partial^\alpha \psi(z')| \leq L \|Wz - Wz'\|^{s_2} d^{\frac{s_1}{2}} \|W\|^{s_1} \leq Ld^{\frac{s_1}{2}} \|W\|^s \|z - z'\|^{s_2}.$$

□

Bounding $\dim(\mathcal{F}(\hat{A}, l, k, R))$

Lemma 25. *We have* $|\Delta_l(R)| \leq \lceil (2^{l+1}R)^d \rceil$*, and thus*

$$\dim(\mathcal{F}(\hat{A}, l, k, R)) \leq \binom{d+k}{k} \lceil (2^{l+1}R)^d \rceil.$$

Proof. First we note that the number of cells with side length 2^{-l} required to cover $[-R, R]^d$ is given by $\lceil (2R2^l)^d \rceil = \lceil (2^{l+1}R)^d \rceil$. Furthermore, for any $w \in \{\hat{A}^\top z : z \in B_R(0)\}$, we have $\|w\| = \left\|\hat{A}^\top z\right\| \leq \|z\| \leq R$, hence $w \in B_R(0)$ (in $\mathbb{R}^d$). Therefore a bound for $|\Delta_l(R)|$ is given by a bound for the number of cells covering $[-R, R]^d$. □

Bounding the approximation error We first show the existence of g^* almost as regular as g and satisfying $g^*(\hat{A}^\top x) \approx g(A^\top x)$. Then we bound the approximation error between f and h over B_R. Finally, we provide the bound for the mean squared approximation error (second term in (24)).

Lemma 26. *Let* g *be* (L, s)*-smooth with* $s = s_1 + s_2$*,* $s_1 \in \mathbb{N}_0$*,* $s_2 \in (0, 1]$*, and* $\|\hat{P} - P\| < 1$*. Then* $(\hat{A}^\top A)^{-1}$ *exists, and the function* $g^*(z) := g((\hat{A}^\top A)^{-1} z)$ *is* (L^*, s)*-smooth for* $L^* := Ld^{s_1/2}(1 - \|\hat{P} - P\|^2)^{-s/2}$*. Moreover, it achieves*

$$\left| g^*(\hat{A}^\top x) - g(A^\top x) \right| \leq L^* \|x\|^{1 \wedge s} \|\hat{P} - P\|^{1 \wedge s}.$$

Proof. Let $\delta := \|\hat{P} - P\| < 1$, and denote the singular value decomposition $\hat{A}^\top A = USV^\top$, where S denotes the cosines of principal angles between $\mathrm{Im}(\hat{A})$ and $\mathrm{Im}(A)$ in descending order. It is known from [20, Definition 2 and Equation (5)] that $\delta = (1 - S_{dd}^2)^{1/2}$, which implies $1 \geq \|S\| \geq \sqrt{1 - \delta^2}$, hence $\hat{A}^\top A$ is invertible with $\|(\hat{A}^\top A)^{-1}\| \leq (1 - \delta^2)^{-1/2}$. Applying Lemma 24, g^* is (L^*, s)-smooth. Furthermore we have $g^*(\hat{A}^\top A A^\top x) = g((\hat{A}^\top A)^{-1} \hat{A}^\top A A^\top x) = g(A^\top x)$. Using the smoothness of g^*, it follows that

$$\begin{aligned}
\left| g^*(\hat{A}^\top x) - g(A^\top x) \right| &= \left| g^*(\hat{A}^\top x) - g^*(\hat{A}^\top A A^\top x) \right| \leq L^* \left\| \hat{A}^\top (\mathsf{Id}_D - P)\, x \right\|^{1 \wedge s} \\
&= L^* \left\| \hat{A}^\top Q x \right\|^{1 \wedge s} \leq L^* \left\| \hat{P} Q \right\|^{1 \wedge s} \|x\|^{1 \wedge s} \\
&= L^* \left\| \hat{P} - P \right\|^{1 \wedge s} \|x\|^{1 \wedge s},
\end{aligned}$$

where we used Lemma 21 in the last equality. □

Proposition 27. *Let* $f(x) = g(A^\top x)$ *for* $P = AA^\top$*,* g *be* (L, s)*-smooth with* $s = s_1 + s_2$*,* $s_1 \in \mathbb{N}_0$*,* $s_2 \in (0, 1]$*, and* $\|\hat{P} - P\| < 1$*. There exists a function* $h \in \mathcal{F}(\hat{A}, l, s_1, R)$ *such that*

$$\max_{x \in B_R} |h(x) - f(x)| \leq L^* \frac{d^{s_1 + \frac{s_2}{2}}}{s_1!} 2^{-s(l+1)} + L^* R^{1 \wedge s} \left\| \hat{P} - P \right\|^{1 \wedge s}, \tag{30}$$

where $L^* := Ld^{s_1/2}(1 - \|\hat{P} - P\|^2)^{-s/2}$.

Proof. First notice that $|h(x) - f(x)| \leq |h(x) - f^*(x)| + |f^*(x) - f(x)|$, where $f^*(x) := g^*(\hat{A}^\top x)$ is the function defined in Lemma 26. Using the bound in Lemma 26, and $\|x\| \leq R$, the second term is bounded by $L^* R^{1\wedge s} \|\hat{P} - P\|^{1\wedge s}$. It remains to bound $|h(x) - f^*(x)|$ for a suitably chosen h. Since $f^*(x) = g^*(\hat{A}^\top x)$ and g^* is (L^*, s)-smooth, we can use the multivariate Taylor theorem to expand g^* as

$$g^*(z) = \sum_{|\alpha| \leq s_1 - 1} \frac{\partial^\alpha g^*(z_0)}{\alpha!}(z - z_0)^\alpha + \sum_{|\alpha| = s_1} \frac{\partial^\alpha g^*(z_0)}{\alpha!}\eta^\alpha \tag{31}$$

for some η on the line segment from z to z_0. We define the function h as follows: for a cell $c \in \Delta_l$, let $z_c \in \mathbb{R}^d$ denote the center point of the cell, and set h_c to

$$h_c(z) := \sum_{|\alpha| \leq s_1} \frac{\partial^\alpha g^*(z_c)}{\alpha!}(z - z_c)^\alpha.$$

Then we define $h \in \mathcal{F}(\hat{A}, l, s_1, R)$ by

$$h(x) := 1_{B_R(0)}(x) \sum_{c \in \Delta_l(R)} 1_c(\hat{A}^\top x) h_c(\hat{A}^\top x) = 1_{B_R(0)}(x) h_{c(x)}(\hat{A}^\top x),$$

where $c(x) := \{c \in \Delta_l(R) : x \in c\}$. To prove (30), we now use (31) with $z_0 = z_{c(x)}$ and compute

$$\begin{aligned} h(x) - g^*(\hat{A}^\top x) &= \sum_{|\alpha| \leq s_1} \frac{\partial^\alpha g^*(z_{c(x)})}{\alpha!} \left(\hat{A}^\top x - z_{c(x)}\right)^\alpha - g^*(\hat{A}^\top x) \\ &= \sum_{|\alpha| = s_1} \frac{\partial^\alpha g^*(z_{c(x)}) - \partial^\alpha g^*(\eta)}{\alpha!} \left(\hat{A}^\top x - z_{c(x)}\right)^\alpha, \end{aligned}$$

where η lies on the line between $\hat{A}^\top x$ and $z_{c(x)}$. The smoothness of g^* implies

$$\begin{aligned} \left|h(x) - g^*(\hat{A}^\top x)\right| &\leq \sum_{|\alpha| = s_1} \frac{\left|\partial^\alpha g^*(z_{c(x)}) - \partial^\alpha g^*(\eta)\right|}{\alpha!} \left|\left(\hat{A}^\top x - z_{c(x)}\right)^\alpha\right| \\ &\leq \sum_{|\alpha| = s_1} \frac{L^* \left\|z_{c(x)} - \eta\right\|^{s_2}}{\alpha!} \left|\left(\hat{A}^\top x - z_{c(x)}\right)^\alpha\right|. \end{aligned}$$

Since $\hat{A}^\top x, z_{c(x)} \in c(x)$, we can furthermore bound

$$\left|\left(\hat{A}^\top x - z_{c(x)}\right)^\alpha\right| = \left|\prod_{i=1}^d ((\hat{A}^\top x)_i - (z_{c(x)})_i)^{\alpha_i}\right| \leq \prod_{i=1}^d \left(2^{-(l+1)}\right)^{\alpha_i} = 2^{-(l+1)s_1}.$$

Furthermore since $c(x)$ is convex, and η is on the line between $A^\top x$ and $c(x)$,

it follows that $\eta \in c(x)$ and therefore also $\left\|z_{c(x)} - \eta\right\| \leq 2^{-(l+1)}\sqrt{d}$. Thus

$$\begin{aligned}\left|h(x) - g^*(\hat{A}^\top x)\right| &\leq L^* d^{\frac{s_2}{2}} 2^{-(l+1)s} \sum_{|\alpha|=s_1} \frac{1}{\alpha!} \\ &= L^* d^{\frac{s_2}{2}} 2^{-(l+1)s} \frac{d^{s_1}}{s_1!} = L^* \frac{d^{s_1+\frac{s_2}{2}}}{s_1!} 2^{-s(l+1)},\end{aligned}$$

where we used the multinomial formula in the second to last equality. □

Corollary 28. *In the setting of Theorem 14, there exists $h \in \mathcal{F}(\hat{A}, l, s_1, R)$ with*

$$\mathbb{E}\left(h(X) - f(X)\right)^2 \leq C_1 N^{-\frac{2s}{2s+d}} + C_2 \log^{1\wedge s}(N) \left\|P - \hat{P}\right\|^{2\wedge 2s},$$

where C_1 depends on L^*, d, s and C_2 depends on L^* and linearly on $(D\left\|X\right\|_{\psi_2}^2)^{1\wedge s}$.

Proof. Using the law of total expectation, and $|h(X) - f(X)| = |f(X)| \leq 1$ if $\|X\| > R$, we obtain for any $h \in \mathcal{F}(\hat{A}, l, s_1, R)$

$$\begin{aligned}\mathbb{E}\left(h(X) - f(X)\right)^2 &\leq \mathbb{E}\left[\left(h(X) - f(X)\right)^2 \,\middle|\, \|X\| \leq R\right] \mathbb{P}(\|X\| \leq R) \\ &\quad + \mathbb{E}\left[\left(h(X) - f(X)\right)^2 \,\middle|\, \|X\| > R\right] \mathbb{P}(\|X\| > R) \\ &\leq \mathbb{E}\left[\left(h(X) - f(X)\right)^2 \,\middle|\, \|X\| \leq R\right] + \mathbb{P}(\|X\| > R).\end{aligned} \tag{32}$$

For the first term, we use the function h in Proposition 27 satisfying the guarantee (30). Using $l = \lceil \log_2(N)/(2s+d) \rceil$, or $2^{-l} \geq N^{-1/(2s+d)}$, and $R^2 = D\left\|X\right\|_{\psi_2}^2 \log(N)$ we get

$$\begin{aligned}&\mathbb{E}\left[\left(h(X) - f(X)\right)^2 \,\middle|\, \|X\| \leq R\right] \\ &\leq \left(L^* \frac{d^{s_1+\frac{s_2}{2}}}{s_1!} 2^{-s(l+1)} + L^* R^{1\wedge s} \left\|\hat{P} - P\right\|^{1\wedge s}\right)^2 \\ &\leq \tilde{C}_1 2^{-2sl} + 2(L^*)^2 R^{2\wedge 2s} \left\|\hat{P} - P\right\|^{2\wedge 2s} \\ &\leq \tilde{C}_1 N^{-\frac{2s}{2s+d}} + 2(L^*)^2 \left(D\left\|X\right\|_{\psi_2}^2 \log(N)\right)^{1\wedge s} \left\|\hat{P} - P\right\|^{2\wedge 2s}.\end{aligned}$$

For the second term in (32), we note that $\|X\|$ is a sub-Gaussian with $\left\|\|X\|\right\|_{\psi_2} \leq \sqrt{D}\left\|X\right\|_{\psi_2}$ by Lemma 19. Therefore, using $R^2 = D\left\|X\right\|_{\psi_2}^2 \log(N)$ we have by [54, Proposition 2.5.2]

$$P(\|X\| > R) \leq \exp\left(-\frac{R^2}{D\left\|X\right\|_{\psi_2}^2}\right) \leq \exp(-\log(N)) = N^{-1}.$$ □

Finalizing the argument

Proof of Theorem 14. Theorem 17 and Corollary 28 imply

$$\begin{aligned}\mathbb{E}\left(\hat{f}(X)-f(X)\right)^2 &\leq C\max\{\sigma_\zeta^2,1\}\frac{\log(N)+\dim(\mathcal{F})}{N}\\ &\quad +C_1' N^{-\frac{2s}{2s+d}}+C_2'\log^{1\wedge s}(N)\left\|\hat{P}-P\right\|^{2\wedge 2s},\end{aligned} \tag{33}$$

where $C_i' = CC_i$ with C_i as in Corollary 28, and C is a universal constant. Furthermore, using Lemma 25, $2^l \leq N^{1/(2s+d)}+1$ and $R^2 = D\|X\|_{\psi_2}^2\log(N)$, we bound the complexity of $\mathcal{F}$ by

$$\begin{aligned}\dim(\mathcal{F}(\hat{A},l,s_1,R)) &\leq \binom{d+s_1}{s_1}\lceil(2^{l+1}R)^d\rceil \leq 2^d\binom{d+s_1}{s_1}\lceil(2^lR)^d\rceil\\ &\leq \binom{d+s_1}{s_1}2^d\left\lceil N^{\frac{d}{2s+d}}\left(D\|X\|_{\psi_2}^2\log(N)\right)^{\frac{d}{2}}\right\rceil\\ &\leq C_3'\log^{d/2}(N)N^{\frac{d}{2s+d}},\end{aligned}$$

with C_3' depending on d, s_1 and linearly on $(D\|X\|_{\psi_2}^2)^{d/2}$. Inserting this in (33) and using $N^{\frac{d}{2s+d}-1} = N^{-\frac{2s}{2s+d}}$, the result follows for $C_2 = C_2'$ and $C_1 = \max\{C_1', C_3', C\max\{\sigma_\zeta^2,1\}\}$. □

Acknowledgements

Timo Klock: this work has been carried out at Simula Research Laboratory (Oslo) and has been supported by the Norwegian Research Council Grant No 251149/O70.
Alessandro Lanteri: this work has been supported by the de Castro Statistics Initiative, Collegio Carlo Alberto, Torino (IT).
Stefano Vigogna: part of this work has been carried out at the Machine Learning Genoa (MaLGa) center, Università di Genova (IT), and has been supported by the European Research Council (grant SLING 819789) and the AFOSR projects FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development).

References

[1] Adragni, K. P. and Cook, R. D. (2009). Sufficient Dimension Reduction and Prediction in Regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367** 4385–4405. MR2546393
[2] Arbel, J., Marchal, O. and Nguyen, H. D. (2020). On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables. *ESAIM: Probability and Statistics* **24** 39–55. MR4053001

[3] Bentley, J. L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* **18** 509–517.
[4] Beygelzimer, A., Kakade, S. and Langford, J. (2006). Cover Trees for Nearest Neighbor. In *Proceedings of the 23rd International Conference on Machine Learning. ICML '06* 97–104.
[5] Bhatia, R. (2013). *Matrix analysis* **169**. Springer Science & Business Media. MR1477662
[6] Bickel, P. J. and Li, B. (2007). Local Polynomial Regression on Unknown Manifolds. *Lecture Notes-Monograph Series* **54** 177–186. MR2459188
[7] Binev, P., Cohen, A., Dahmen, W. and DeVore, R. (2007). Universal Algorithms for Learning Theory. Part II: Piecewise Polynomial Functions. *Constructive Approximation* **26** 127–152. MR2327596
[8] Boente Boente, G. L. and Rodriguez, D. A. (2012). Robust Estimates in Generalized Partially Linear Single-Index Models. **21** 386–441. MR2935366
[9] Chen, D., Hall, P., Müller, H.-G. et al. (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics* **39** 1720–1747. MR2850218
[10] Cook, R. D. (1994). On the Interpretation of Regression Plots. *Journal of the American Statistical Association* **89** 177–189. MR1266295
[11] Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics* **482**. John Wiley & Sons. MR1645673
[12] Cook, R. D. (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods* **29** 2109–2121.
[13] Cook, R. D. and Li, B. (2002). Dimension Reduction for Conditional Mean in Regression. *The Annals of Statistics* **30** 455–474. MR1902895
[14] Cook, R. D. and Li, B. (2004). Determining the dimension of iterative Hessian transformation. *The Annals of Statistics* **32** 2501–2531. MR2153993
[15] Dalalyan, A. S., Juditsky, A. and Spokoiny, V. (2008). A New Algorithm for Estimating the Effective Dimension-Reduction Subspace. *Journal of Machine Learning Research* **9** 1647–1678. MR2438819
[16] Fornasier, M., Schnass, K. and Vybíral, J. (2012). Learning Functions of Few Arbitrary Linear Parameters in High Dimensions. *Foundations of Computational Mathematics* **12** 229–262. MR2898783
[17] Fornasier, M., Vybíral, J. and Daubechies, I. (2018). Identification of Shallow Neural Networks by Fewest Samples. *ArXiv preprint arXiv:1804.01592*.
[18] Guo, X., Xu, W. and Zhu, L. (2014). Multi-index regression models with missing covariates at random. *Journal of Multivariate Analysis* **123** 345–363. MR3130439
[19] Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2006). *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media. MR1920390
[20] Hamm, J. and Lee, D. D. (2008). Grassmann discriminant analysis: a

unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning* 376–383. ACM.

[21] Hardle, W. and Stoker, T. M. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association* **84** 986–995. MR1134488

[22] Hemant, T. and Cevher, V. (2012). Active Learning of Multi-Index Function Models. In *Advances in NeurIPS* 1466–1474.

[23] Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001). Structure Adaptive Approach for Dimension Reduction. *The Annals of Statistics* **29** 1537–1566. MR1891738

[24] Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics* 595–623. MR1865333

[25] Janzamin, M., Sedghi, H. and Anandkumar, A. (2015). Beating the Perils of Non-Convexity: Guaranteed Training of Neural Networks using Tensor Methods. *ArXiv preprint arXiv:1506.08473*.

[26] Kato, T. (2013). *Perturbation Theory for Linear Operators* **132**. Springer Science & Business Media. MR0203473

[27] Kereta, v. and Klock, T. (2021). Estimating covariance and precision matrices along subspaces. *Electronic Journal of Statistics* **15** 554–588. MR4198545

[28] Kereta, Z., Klock, T. and Naumova, V. (2020). Nonlinear generalization of the monotone single index model. *Information and Inference: A Journal of the IMA*.

[29] Kohler, M., Krzyżak, A. and Walk, H. (2006). Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *Journal of Multivariate Analysis* **97** 311–323. MR2234025

[30] Kpotufe, S. (2011). k-NN Regression Adapts to Local Intrinsic Dimension. *Advances in Neural Information Processing Systems 24* 729–737.

[31] Kpotufe, S. and Garg, V. (2013). Adaptivity to Local Smoothness and Dimension in Kernel Regression. *Advances in Neural Information Processing Systems 26* 3075–3083.

[32] Kuchibhotla, A. K. and Patra, R. K. (2016). Efficient Estimation in Single Index Models through Smoothing Splines. *ArXiv preprint arXiv:1612.00068*. MR4058379

[33] Lanteri, A., Maggioni, M. and Vigogna, S. (2020). Conditional regression for single-index models. *ArXiv preprint arXiv:2002.10008*.

[34] Larsson, E. G. and Selén, Y. (2007). Linear Regression With a Sparse Parameter Vector. *IEEE Transactions on Signal Processing* **55** 451–460. MR2445956

[35] Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R. Chapman & Hall/CRC Monographs on Statistics and Applied Probability*. CRC Press.

[36] Li, B. and Wang, S. (2007). On Directional Regression for Dimension Reduction. *Journal of the American Statistical Association* **102** 997–1008. MR2354409

[37] Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics* **33** 1580–1616. MR2166556
[38] Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* **86** 316–327. MR1137117
[39] Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction: Rejoinder. *Journal of the American Statistical Association* **86** 337–342. MR1137117
[40] Li, K.-C. (1992). On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association* **87** 1025–1039. MR1209564
[41] Liao, W., Maggioni, M. and Vigogna, S. (2016). Learning adaptive multiscale approximations to data and functions near low-dimensional sets. In *2016 IEEE Information Theory Workshop (ITW)* 226–230. IEEE.
[42] Liao, W., Maggioni, M. and Vigogna, S. (2021). Multiscale regression on unknown manifolds. *ArXiv e-prints arXiv:2101.05119*.
[43] Liu, J., Zhang, R., Zhao, W. and Lv, Y. (2013). A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis* **122** 226–238. MR3189320
[44] Ma, Y. and Zhu, L. (2012). A Semiparametric Approach to Dimension Reduction. *Journal of the American Statistical Association* **107** 168–179. MR2949349
[45] Ma, Y. and Zhu, L. (2013). A Review on Dimension Reduction. *International Statistical Review* **81** 134–150. MR3047506
[46] Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics* **41** 250. MR3059417
[47] Mitzenmacher, M. and Upfal, E. (2017). *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press. MR3674428
[48] Mondelli, M. and Montanari, A. (2018). On the Connection Between Learning Two-Layers Neural Networks and Tensor Decomposition. *ArXiv preprint arXiv:1802.07301*.
[49] Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis* **139** 266–282. MR3349492
[50] Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57** 6976–6994. MR2882274
[51] Reiss, M., Wahl, M. et al. (2020). Nonasymptotic upper bounds for the reconstruction error of PCA. *Annals of Statistics* **48** 1098–1123. MR4102689
[52] Soudry, D. and Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks. *ArXiv preprint arXiv:1605.08361*.
[53] Stone, C. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics* **10** 1040–1053. MR0673642
[54] Vershynin, R. (2018). *High-dimensional probability: An introduction with*

applications in data science **47**. Cambridge University Press. MR3837109

[55] Wang, H. and Xia, Y. (2008). Sliced Regression for Dimension Reduction. *Journal of the American Statistical Association* **103** 811–821. MR2524332

[56] Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen* **71** 441–479. MR1511670

[57] Xia, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association* **103** 1631–1640. MR2504209

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

[59] Zhu, L.-X., Ohtaki, M. and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics & Data Analysis* **51** 2621–2635. MR2338992