Supplements to "Gradient-based kernel dimension reduction for feature extraction and variable selection"

A Table 1 with standard errors

The following is the experimental results with standard errors for Table 1. See the main body of the paper for the experimental setting. The large standard errors in the KDR show that the optimization sometimes fails to find the optimal projector.

	gKDR -FEX	gKDR -FEXi	gKDR -FEXv	IADE	SIR II	KDR	gKDR-FEX +KDR
(A) $n = 100$	0.1989	0.1639	0.2002	0.1372	0.2986	0.2807	0.0883
	(0.0553)	(0.0479)	(0.0555)	(0.0552)	(0.1021)	(0.3364)	(0.1473)
(A) $n = 200$	0.1264	0.0995	0.1287	0.0857	0.2077	0.1175	0.0501
	(0.0321)	(0.0352)	(0.0351)	(0.0258)	(0.0554)	(0.2184)	(0.0964)
(B) $n = 100$	0.1500	0.1358	0.1630	0.1690	0.3137	0.2138	0.1076
	(0.0363)	(0.0331)	(0.0325)	(0.0624)	(0.0679)	(0.2202)	(0.0967)
(B) $n = 200$	0.0755	0.0750	0.0802	0.0940	0.2129	0.1440	0.0506
	(0.0157)	(0.0153)	(0.0160)	(0.0318)	(0.0359)	(0.2190)	(0.0729)
(C) $n = 200$	0.1919	0.2322	0.1930	0.7724	0.7326	0.1479	0.1285
	(0.0791)	(0.1512)	(0.0763)	(0.1665)	(0.0153)	(0.1307)	(0.0483)
(C) $n = 400$	0.1346	0.1372	0.1369	0.7863	0.7167	0.0897	0.0893
	(0.0472)	(0.0644)	(0.0499)	(0.1846)	(0.0470)	(0.0294)	(0.0294)

Table 5: gKDE-FEX for synthetic data: mean discrepancies and standard errors (in brackets) over 100 runs.

B Consistency of the kernel estimator for the regression function

We discuss the consistency of the estimator $\left(\widehat{C}_{XX} + \varepsilon_n I\right)^{-1} \widehat{C}^{(n)}_{XY} g$ for $E[g(Y)|X = \cdot]$. While this consistency has been already proved in some literature such as [2, 4, 8, 9] in various contexts, we show the proof in our terminology for completeness.

Theorem 5. *Let $g \in \mathcal{H}_\mathcal{Y}$ and assume that $E[g(Y)|X = \cdot] \in \mathcal{R}(C^\nu_{XX})$ for $\nu \geq 0$, where $\mathcal{R}(C^0_{XX})$ for $\nu = 0$ is interpreted as $\mathcal{H}_\mathcal{X}$. If $\varepsilon_n \to 0$ ($n \to \infty$), then*

$$\left\| \left(\widehat{C}^{(n)}_{XX} + \varepsilon_n I\right)^{-1} \widehat{C}^{(n)}_{XY} g - E[g(Y)|X = \cdot] \right\|_{\mathcal{H}_\mathcal{X}}$$

is of the order

$$\begin{cases} O_p(\varepsilon_n^{-1} n^{-1/2}) + O(\varepsilon_n^\nu), & \text{for } 0 \leq \nu < 1, \\ O_p(\varepsilon_n^{-1} n^{-1/2}) + O(\varepsilon_n), & \text{for } \nu \geq 1. \end{cases}$$

Consequently, if $\varepsilon_n = n^{-\max\{\frac{1}{4}, \frac{1}{2\nu+2}\}}$, then the estimator is consistent of the order $O\left(n^{-\min\{\frac{1}{4}, \frac{\nu}{2\nu+2}\}}\right)$.

Proof. Take $\eta \in \mathcal{H}_\mathcal{X}$ such that $E[g(Y)|X = \cdot] = C^\nu_{XX}\eta$. From the fact $C_{XX}E[g(Y)|X = \cdot] = C_{XY}g$ ([5], Theorem 2), we have $C_{XY}g = C_{XX}E[g(Y)|X = \cdot] = C^{\nu+1}_{XX}\eta$.

First, we show

$$\left\| \left(\widehat{C}^{(n)}_{XX} + \varepsilon_n I\right)^{-1} \widehat{C}^{(n)}_{XY} g - (C_{XX} + \varepsilon_n I)^{-1} C_{XY} g \right\|_{\mathcal{H}_\mathcal{X}} = O_p(\varepsilon_n^{-1} n^{-1/2}) \qquad (n \to \infty). \quad (11)$$

Since $B^{-1} - A^{-1} = B^{-1}(A - B)A^{-1}$ for any invertible operators A and B, the left hand side is upper bounded by

$$\left\| \left(\widehat{C}^{(n)}_{XX} + \varepsilon_n I\right)^{-1} (C_{XX} - \widehat{C}^{(n)}_{XX})(C_{XX} + \varepsilon_n I)^{-1} C_{XY} g \right\|_{\mathcal{H}_\mathcal{X}} + \left\| \left(\widehat{C}^{(n)}_{XX} + \varepsilon_n I\right)^{-1} (\widehat{C}^{(n)}_{XY} - C_{XY}) g \right\|_{\mathcal{H}_\mathcal{X}}.$$

From $C_{XY}g = C_{XX}^{\nu+1}\eta$, we have $\|(C_{XX}+\varepsilon_n I)^{-1}C_{XY}g\| \leq \|C_{XX}^{\nu}\eta\|_{\mathcal{H}_\mathcal{X}}$. Combination of this fact with $\|\widehat{C}_{XX}^{(n)} - C_{XX}\| = O_p(n^{-1/2})$ proves that the first term is of the order $O_p(\varepsilon_n^{-1}n^{-1/2})$. The second term is of the same order from $\|\widehat{C}_{XY}^{(n)} - C_{XY}\| = O_p(n^{-1/2})$, which implies Eq. (11).

Next, we derive the upper bounds

$$\left\|\left(C_{XX}+\varepsilon_n I\right)^{-1}C_{XY}g - E[g(Y)|X=\cdot]\right\|_{\mathcal{H}_\mathcal{X}} = \begin{cases} O(\varepsilon_n^{\nu}), & \text{for } 0 \leq \nu < 1, \\ O(\varepsilon_n), & \text{for } \nu \geq 1. \end{cases} \tag{12}$$

It follows from $E[g(Y)|X=\cdot] = C_{XX}^{\nu}\eta$ and $C_{XY}g = C_{XX}^{\nu+1}\eta$ that

$$(C_{XX}+\varepsilon_n I)^{-1}C_{XY}g - E[g(Y)|X=\cdot] = (C_{XX}+\varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}\eta.$$

Let $C_{XX} = \sum_i \lambda_i \phi_i \langle \phi_i, \cdot \rangle$ be the eigendecomposition of C_{XX} such that $\lambda_i > 0$ are the eigenvalues and ϕ_i are the ohorthonormal eigenvectors. The eigendespectrum of the operator $(C_{XX}+\varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}$ is then given by

$$\frac{\lambda_i^{\nu+1}}{\lambda_i+\varepsilon_n} - \lambda_i^{\nu} = \frac{\lambda_i^{\nu}\varepsilon_n}{\lambda_i+\varepsilon_n} \qquad (i=1,2,\ldots).$$

If $0 \leq \nu < 1$, from $\frac{\lambda^{\nu}\varepsilon_n}{\lambda+\varepsilon_n} = \varepsilon_n^{\nu}\frac{\lambda^{\nu}\varepsilon_n^{1-\nu}}{\lambda+\varepsilon_n} \leq \varepsilon_n^{\nu}\frac{\varepsilon_n^{1-\nu}}{(\lambda+\varepsilon_n)^{1-\nu}}$ and $\left|\frac{\varepsilon_n^{1-\nu}}{(\lambda+\varepsilon_n)^{1-\nu}}\right| \leq 1$ we have

$$\|(C_{XX}+\varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}\| \leq \varepsilon_n^{\nu}.$$

If $\nu \geq 1$, then $\frac{\lambda^{\nu}\varepsilon_n}{\lambda+\varepsilon_n} \leq \varepsilon_n\frac{\lambda^{\nu}}{\lambda+\varepsilon_n} \leq \varepsilon_n\lambda^{\nu-1}$. It follows

$$\|(C_{XX}+\varepsilon_n I)^{-1}C_{XX}^{\nu+1}\eta - C_{XX}^{\nu}\| \leq \varepsilon_n\|C_{XX}\|^{\nu-1}.$$

From Eqs. (11) and (12), the proof is completed. □

C Proof of Theorems in Section 2.4

Proof of Theorem 1. Note that, from Eqs. (6) and (7), the eigenvectors of $E[M(x)]$ is contained in $\mathrm{Span}(B)$ if and only if $\partial E[g(Y)|X=x]/\partial x \in \mathrm{Span}(B)$ for any $g \in \mathcal{H}_\mathcal{Y}$ almost surely.

Let $B_\perp$ be an $m \times (m-d)$ matrix such that $B_\perp^T B_\perp = I_{m-d}$ and the column vectors of $B_\perp$ are orthogonal to those of B, and write $(U,V) = (B^T X, B_\perp^T X)$. Then, the condition $\partial E[g(Y)|X=x]/\partial x \in \mathrm{Span}(B)$ almost surely is equivalent to $E[g(Y)|(U,V)=(u,v)] = E[g(Y)|U=u]$ for any $g \in \mathcal{H}_\mathcal{Y}$ almost surely. Since $k_\mathcal{Y}$ is characteristic, this implies that the conditional probability of Y given (U,V) is equal to that of Y given U, which means the desired conditional independence. □

Proof of Theorem 2. Let $g_a = \frac{\partial k_\mathcal{X}(\cdot,x)}{\partial x^a}$. Since

$$\begin{aligned} M_{ab}(x) &= \left\langle \left\langle E[k_\mathcal{Y}(*,Y)|X=\cdot], g_a \right\rangle_{\mathcal{H}_\mathcal{X}}, \left\langle E[k_\mathcal{Y}(*,Y)|X=\cdot], g_b \right\rangle_{\mathcal{H}_\mathcal{X}} \right\rangle_{\mathcal{H}_\mathcal{Y}} \\ &= \left\langle E[k_\mathcal{Y}(*,Y)|g_a(X)], E[k_\mathcal{Y}(*,Y)|g_b(X)] \right\rangle_{\mathcal{H}_\mathcal{Y}} \end{aligned}$$

and

$$\widehat{M}_{n,ab}(x) = \left\langle \widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1} g_a, \widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1} g_b \right\rangle_{\mathcal{H}_\mathcal{Y}},$$

we have

$$\begin{aligned} &\left|\widehat{M}_{n,ab}(x) - M_{ab}(x)\right| \\ &\leq \left|\left\langle \widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1} g_a, \widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1} g_b - E[k_\mathcal{Y}(*,Y)|g_b(X)] \right\rangle_{\mathcal{H}_\mathcal{Y}}\right| \\ &+ \left|\left\langle \widehat{C}_{YX}^{(n)}\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1} g_a - E[k_\mathcal{Y}(*,Y)|g_a(X)], E[k_\mathcal{Y}(*,Y)|g_b(X)] \right\rangle_{\mathcal{H}_\mathcal{Y}}\right|. \end{aligned}$$

Noting $\varepsilon_n\sqrt{n} \to \infty$ and the expression

$$\left(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\right)^{-1} = (C_{XX}+\varepsilon_n I)^{-1}\left\{I - \left(C_{XX} - \widehat{C}_{XX}^{(n)}\right)(C_{XX}+\varepsilon_n I)^{-1}\right\}^{-1},$$

Lemma 4 in [9] shows $\|\big(C_{XX}-\widehat{C}_{XX}^{(n)}\big)(C_{XX}+\varepsilon_n I)^{-1}\|_{HS} = O_p(\varepsilon_n^{-1}n^{-1/2})$. Noting $\varepsilon_n\sqrt{n}\to\infty$ and the expression

$$\big(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\big)^{-1} = (C_{XX}+\varepsilon_n I)^{-1}\big\{I-\big(C_{XX}-\widehat{C}_{XX}^{(n)}\big)(C_{XX}+\varepsilon_n I)^{-1}\big\}^{-1},$$

we obtain

$$\big\|C_{XX}\big(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\big)^{-1}\big\| = O_p(1).$$

From $g_a = C_{XX}^{\beta+1}\eta$ for some $\eta\in\mathcal{H}_{\mathcal{X}}$, we have $\|\widehat{C}_{YX}^{(n)}\big(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\big)^{-1}g_a\| = O_p(1)$. For the proof of the first assertion of Theorem 2, it is then sufficient to prove the following theorem.

Theorem 6. *Assume that $g\in\mathcal{H}_{\mathcal{X}}$ satisfies $\mathcal{R}(C_{XX}^{\beta+1})$ for some $\beta\geq 0$ and that $E[k_{\mathcal{Y}}(y,Y)|X=\cdot]\in\mathcal{H}_{\mathcal{X}}$ for every $y\in\mathcal{Y}$. Then, for $\varepsilon_n>0$ with $\varepsilon_n = n^{-\max\{\frac{1}{3},\frac{1}{2(\beta+1)}\}}$, we have*

$$\big\|\widehat{C}_{YX}^{(n)}\big(\widehat{C}_{XX}^{(n)}+\varepsilon_n I_n\big)^{-1}g - E[k_{\mathcal{Y}}(\cdot,Y)|g(X)]\big\|_{\mathcal{H}_{\mathcal{Y}}} = O_p\Big(n^{-\min\{\frac{1}{3},\frac{2\beta+1}{4\beta+4}\}}\Big)$$

as $n\to\infty$.

Proof. It suffices to show

$$\big\|\widehat{C}_{YX}^{(n)}\big(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\big)^{-1}g - C_{YX}\big(C_{XX}+\varepsilon_n I\big)^{-1}g\big\|_{\mathcal{H}_{\mathcal{Y}}}^2 = O_p\big(\varepsilon_n^{-1/2}n^{-1/2}\big) \tag{13}$$

and

$$\big\|C_{YX}\big(C_{XX}+\varepsilon_n I\big)^{-1}g - E[k_{\mathcal{Y}}(\cdot,Y)|g(X)]\big\|_{\mathcal{H}_{\mathcal{Y}}}^2 = O\big(\varepsilon_n^{\min\{1,(2\beta+1)/2\}}\big) \tag{14}$$

as $n\to\infty$. In fact, optimizing the rate derives the assertion of the theorem.

Let $g = C_{XX}^{\beta+1}h$, where $h\in\mathcal{H}_{\mathcal{X}}$. Since $B^{-1}-A^{-1} = B^{-1}(A-B)A^{-1}$ for any invertible operators A and B, the left hand side of Eq. (13) is upper bounded by

$$\begin{aligned}\big\|\widehat{C}_{YX}^{(n)}\big(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\big)^{-1}(C_{XX}-\widehat{C}_{XX}^{(n)})\big(C_{XX}+\varepsilon_n I\big)^{-1}C_{XX}^{\beta+1}h\big\|_{\mathcal{H}_{\mathcal{Y}}}&\\ + \big\|(\widehat{C}_{YX}^{(n)}-C_{YX})\big(C_{XX}+\varepsilon_n I\big)^{-1}C_{XX}^{\beta+1}h\big\|_{\mathcal{H}_{\mathcal{Y}}}.&\end{aligned}$$

By the decomposition $\widehat{C}_{YX}^{(n)} = \widehat{C}_{YY}^{(n)1/2}\widehat{W}_{YX}\widehat{C}_{XX}^{(n)1/2}$ with $\|\widehat{W}_{YX}\|\leq 1$ ([1]), we have $\|\widehat{C}_{YX}^{(n)}\big(\widehat{C}_{XX}^{(n)}+\varepsilon_n I\big)^{-1}\| = O(\varepsilon_n^{-1/2})$. It is known that $\|C_{XX}-\widehat{C}_{XX}^{(n)}\| = O_p(n^{-1/2})$. From these two fact, we see that the first term is of $O_p(\varepsilon_n^{-1/2}n^{-1/2})$. Since the second term is of $O_p(n^{-1/2})$, Eq. (13) is obtained.

For Eq. (14), first note that for each y

$$\begin{aligned}E[k_{\mathcal{Y}}(y,Y)|g(X)] &= \langle E[k_{\mathcal{Y}}(y,Y)|X=\cdot], g\rangle = \langle E[k_{\mathcal{Y}}(y,Y)|X=\cdot], C_{XX}^{\beta+1}h\rangle\\ &= \langle C_{XX}E[k_{\mathcal{Y}}(y,Y)|X=\cdot], C_{XX}^{\beta}h\rangle = \langle C_{XY}k_{\mathcal{Y}}(y,\cdot), C_{XX}^{\beta}h\rangle\\ &= \langle k_{\mathcal{Y}}(y,\cdot), C_{YX}C_{XX}^{\beta}h\rangle = (C_{YX}C_{XX}^{\beta}h)(y),\end{aligned}$$

which means $E[k_{\mathcal{Y}}(\cdot,Y)|g(X)] = C_{YX}C_{XX}^{\beta}h$. Let $C_{YX} = C_{YY}^{1/2}W_{YX}C_{XX}^{1/2}$ be the decomposition with $\|W_{YX}\|\leq 1$. Then, we have

$$\begin{aligned}\big\|C_{YX}\big(C_{XX}+\varepsilon_n I\big)^{-1}g - E[k_{\mathcal{Y}}(\cdot,Y)|g(X)]\big\|_{\mathcal{H}_{\mathcal{Y}}}&\\ = \|C_{YY}^{1/2}W_{YX}\|\big\|C_{XX}^{\beta+3/2}\big(C_{XX}+\varepsilon_n I\big)^{-1}h - C_{XX}^{\beta+1/2}h\big\|_{\mathcal{H}_{\mathcal{Y}}}.&\end{aligned}$$

Let $\{\phi_i\}$ be the unit eigenvectors of C_{XX} such that $C_{XX}f = \sum_i \lambda_i\langle\phi_i, f\rangle$. Then the eigenspectrum of $C_{XX}^{\beta+3/2}\big(C_{XX}+\varepsilon_n I\big)^{-1} - C_{XX}^{\beta+1/2}$ is given by

$$-\frac{\varepsilon_n\lambda_i^{(2\beta+1)/2}}{\lambda_i+\varepsilon_n}\qquad(i=1,2,\ldots).$$

If $0\leq\beta<1/2$, we have $\frac{\varepsilon_n\lambda_i^{(2\beta+1)/2}}{\lambda_i+\varepsilon_n} = \frac{\lambda_i^{(2\beta+1)/2}}{(\lambda_i+\varepsilon_n)^{(2\beta+1)/2}}\frac{\varepsilon_n^{(1-2\beta)/2}}{(\lambda_i+\varepsilon_n)^{(1-2\beta)/2}}\varepsilon_n^{(2\beta+1)/2}\leq\varepsilon_n^{(2\beta+1)/2}$. If $\beta\geq 1/2$, then $\frac{\varepsilon_n\lambda_i^{(2\beta+1)/2}}{\lambda_i+\varepsilon_n}\leq\lambda_i^{\beta-1/2}\varepsilon_n$. We have thus Eq. (14), which completes the proof of Theorem 6

□

For the second assertion of Theorem 2, note

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \widehat{M}_n(X_i) - E[M(X)] \right\|_F \leq \left\| \frac{1}{n}\sum_{i=1}^{n} \widehat{M}_n(X_i) - \frac{1}{n}\sum_{i=1}^{n} M(X_i) \right\|_F + \left\| \frac{1}{n}\sum_{i=1}^{n} M(X_i) - E[M(X)] \right\|_F .$$

The second term in the right hand side is of $O_p(n^{-1/2})$ by the central limit theorem. By replacing g and h in the proof of Theorem 6 by $\sum_{i=1}^{n} g_{X_i}/n$ and $\sum_{i=1}^{n} h_{X_i}/n$, respectively, the assertion is obtained as a corollary.

D Proof of Theorems in Section 3.2

The proof is essentially the same as that of Theorems 1 and 2 in Chen et al. [3] with appropriate change of the consistency rate, but we show the proofs for completeness.

Let $\mathrm{St}(m,d)$ denote the Stiefel manifold, that is, the space of $m \times d$ matrix B with $B^T B = I_d$. We first summarize fundamental facts on Stiefel manifolds (see [6] for details). For $B \in \mathrm{St}(m,d)$, the tangent space of $\mathrm{St}(m,d)$ at B is denoted by $T_B(m,d)$. It is known that $T_B(m,d)$ is given by

$$T_B(m,d) = \{Z \in \mathbb{R}^{m\times d} \mid Z = BF + B_\perp G, F \in \mathbb{R}^{d\times d}, F + F^T = 0, G \in \mathbb{R}^{(m-d)\times d}\},$$

where $B_\perp$ is a matrix in $\mathrm{St}(m, m-d)$ such that $(B, B_\perp) \in O(m)$.

Let Π denote the projection of a general $m \times d$ matrix onto the Stiefel manifold $\mathrm{St}(m,d)$,

$$\Pi(W) = \arg\min_{B\in\mathrm{St}(m,d)} \|B - W\|_F.$$

Lemma 7. *Let $B \in \mathrm{St}(m,d)$ and $Z \in T_B(m,d)$, then*

$$\Pi(B + tZ) = B + tZ - \frac{t^2}{2} BZ^T Z + O(t^3)$$

as $t \to 0$.

Let $Q(B)$ denote the objective function of gKDR-VS, i.e.,

$$Q(B) = -\mathrm{Tr}[B^T \tilde{M}_n B] + \sum_{j=1}^{m} \lambda_j \|\mathbf{v}_j\|.$$

For $B \in \mathrm{St}(m,d)$, let $[B]$ denote the corresponding element in Gramsmann manifold $\mathrm{Gr}(m,d)$, which is the manifold of d dimensional subspaces in $\mathbb{R}^m$. Note that the first term of $Q(B)$ depends only on $[B]$.

For $Z \in T_B(m,d)$ consider a perturbation $\Pi(B + tZ)$. It is known [3] that if t is sufficiently small, there exists $G \in \mathbb{R}^{(m-d)\times d}$ such that

$$[\Pi(B + tZ)] = [\Pi(B + tB_\perp G)]. \tag{15}$$

We can thus use only the $B_\perp G$ component of the tangent space, when the contribution of perturbation is considered in Gramsmann manifold .

Proof of Theorem 3. It suffices to prove that for any $\varepsilon > 0$ there is $C > 0$ such that

$$\lim_{n\to\infty} \Pr\Big(\inf_{Z\in T_{B_0}(m,d), \|Z\|=C} Q(\Pi(B_0 + n^{-\tau} Z)) > Q(B_0) \Big) > 1 - \varepsilon.$$

Let $e_j \in \mathbb{R}^d$ be the vector with the j-th component 1 and others 0. For $Z \in T_{B_0}(m,d)$, using Lemma 7 we have

$$\begin{aligned}
&n^{2\tau}\big[Q(\Pi(B_0+n^{-\tau}Z))-Q(B_0)\big]\\
\geq&\Big[-\mathrm{Tr}[Z\tilde{M}_nZ]+\mathrm{Tr}[B_0^T\tilde{M}_nB_0Z^TZ]-2n^{\tau}\mathrm{Tr}[B_0^T\tilde{M}_nZ]\Big](1+o_p(1))\\
&+\sum_{j=1}^{q}n^{2\tau}\lambda_j\Big(\Big\|e_j^T\Big(B_0+n^{-\tau}Z-\frac{n^{-2\tau}}{2}B_0Z^TZ\Big)\Big\|-\|\mathbf{v}_{0j}\|\Big)(1+o_p(1))\\
\geq&n^{2\tau}\Big[-\mathrm{Tr}[Z\tilde{M}_nZ]+\mathrm{Tr}[B_0^T\tilde{M}_nB_0Z^TZ]-2n^{\tau}\mathrm{Tr}[B_0^T\tilde{M}_nZ]\Big](1+o_p(1))\\
&-\frac{q\alpha_nn^{\tau}}{2}\max_{1\leq j\leq q}\frac{\|e_j^T(Z-\frac{n^{-\tau}}{2}B_0Z^TZ)\|}{\|\mathbf{v}_{0j}\|}(1+o_p(1))
\end{aligned}\tag{16}$$

where the first inequality holds by $\mathbf{v}_{0j}=0$ for $q+1\leq j\leq m$, and the second one is given by Taylor expansion.

By the assumption $\mathbf{v}_{0j}\neq 0$ for $1\leq j\leq q$ and $\alpha_nn^{\tau}\to 0$ as $n\to\infty$, the second term of Eq. (16) converges to zero in probability.

For the first term, note that from Eq. (15) we can assume that $Z=B_{0\perp}G$. Since B_0 consists of the top d eigenvectors of M, we have $B_0^TMB_0=\Lambda_{(d)}$, where $\Lambda_{(d)}$ is the diagonal matrix with the largest d eigenvalues. We have

$$\begin{aligned}
\big|n^{\tau}\mathrm{Tr}[B_0^T\tilde{M}_nZ]\big|&\leq\big|n^{\tau}\mathrm{Tr}\big[B_0^T(\tilde{M}_n-M)Z\big]\big|+\big|n^{\tau}\mathrm{Tr}\big[B_0^TMZ\big]\big|\\
&\leq\big|n^{\tau}\mathrm{Tr}\big[B_0^T(\tilde{M}_n-M)Z\big]+n^{\tau}\mathrm{Tr}[\Lambda_{(d)}B_0^TB_{0\perp}G]\\
&=\|Z\|_FO_p(1)+0,
\end{aligned}$$

and

$$\begin{aligned}
&-\mathrm{Tr}[Z\tilde{M}_nZ]+\mathrm{Tr}[B_0^T\tilde{M}_nB_0Z^TZ]\\
=&-\mathrm{Tr}[Z^TMZ]+\mathrm{Tr}[B_0^TMB_0Z^TZ]+O_p(n^{-\tau})\\
\geq&-\eta_{d+1}\mathrm{Tr}[Z^TZ]+\mathrm{Tr}[\Lambda_{(d)}Z^TZ]+O_p(n^{-\tau})\\
\geq&(\eta_d-\eta_{d+1})\|Z\|_F^2+O_p(n^{-\tau}).
\end{aligned}$$

It follows from the assumption $\eta_d>\eta_{d+1}$ that Eq. (16) is positive for sufficiently large $\|Z\|$, which completes the proof. □

Proof of Theorem 4. For simplicity we write $\widehat{B}$ for $\widehat{B}_\lambda$ in the proof, and $\widehat{B}=(\widehat{\mathbf{v}}_1^T,\ldots,\widehat{\mathbf{v}}_m^T)^T=(\widehat{\mathbf{b}}_1,\ldots,\widehat{\mathbf{b}}_d)$. The optimization of gKDR-VS is written as

$$\min f(B)+\rho(B)\qquad\text{subject to}\quad\begin{cases}\mathbf{b}_j^T\mathbf{b}_j=1 & (1\leq j\leq m),\\ \mathbf{b}_i^T\mathbf{b}_j=0 & (1\leq i<j\leq m),\end{cases}$$

where $f(B):=-\mathrm{Tr}[B^T\tilde{M}_nB]$ and $\rho(B):=\sum_{j=1}^m\lambda_j\|\mathbf{v}_j\|$.

Suppose $\widehat{\mathbf{v}}_j\neq 0$ for all j. Let $\mathbf{t}=\mathrm{vec}(B)$. It is known [7] that in this case the Laglangian multiplier rule gives

$$\mathcal{R}\frac{\partial f(B)}{\partial\mathbf{t}}\Big|_{\mathbf{t}=\widehat{\mathbf{t}}}+\mathcal{R}\frac{\partial\rho(B)}{\partial\mathbf{t}}\Big|_{\mathbf{t}=\widehat{\mathbf{t}}}=0,\tag{17}$$

where $\mathcal{R}=(I-UU^T)$ is the projection matrix with $U=(\mathbf{u}_{11},\ldots,\mathbf{u}_{1d},\mathbf{u}_{22},\ldots,\mathbf{u}_{2d},\ldots,\mathbf{u}_{d-1,d-1},\mathbf{u}_{d-1,d},\mathbf{u}_{dd})\in\mathbb{R}^{m\times d(d+1)/2}$. The $m\times d$-dimensional

vector $\mathbf{u}_{ij}$ is defined by

$$\mathbf{u}_{ii} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \widehat{\mathbf{b}}_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} i\text{-th} \quad (1 \le i \le d), \qquad \mathbf{u}_{ij} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \widehat{\mathbf{b}}_j \\ 0 \\ \vdots \\ 0 \\ \widehat{\mathbf{b}}_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \\ \\ \\ i\text{-th} \\ \\ \\ \\ j\text{-th} \\ \\ \\ \end{matrix} \quad (1 \le i < j \le d).$$

Note that $\mathbf{u}_{ij}$ ($i < j$) has $\widehat{\mathbf{b}}_j$ at the i-th block and $\widehat{\mathbf{b}}_i$ at the j-th block, and that $U^T U = I_{d(d+1)/2}$.

Let $\tilde{B} = \arg\min_{B \in \mathrm{St}(m,d)} f(B)$, i.e., the top d eigenvectors of $\tilde{M}_n$. Then, $\mathcal{R}\frac{\partial f(B)}{\partial \mathbf{t}}\big|_{B=\tilde{B}} = 0$. Since $\frac{\partial f(B)}{\partial \mathbf{t}}\big|_{B=\tilde{B}}$ is linear with respect to $\mathbf{t}$, and since $D(\tilde{B}, B_0) = O_p(n^{-\tau})$ from perturbation theory and $D(\widehat{B}, B_0) = O_p(n^{-\tau})$ from Theorem 3, we have

$$\mathcal{R}\frac{\partial \rho(B)}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\widehat{\mathbf{t}}} = O_p(n^{-\tau}).$$

By the definition of $\mathcal{R}$, there are $\gamma_{ij} \in \mathbb{R}$ ($1 \le i \le j \le d$) such that

$$\frac{\partial \rho(B)}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\widehat{\mathbf{t}}} = \sum_{i \le j} \gamma_{ij} \mathbf{u}_{ij} + O_p(n^{-\tau}). \tag{18}$$

Note that

$$\frac{\partial \rho(B)}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\widehat{\mathbf{t}}} = \begin{pmatrix} D\widehat{\mathbf{b}}_1 \\ \vdots \\ D\widehat{\mathbf{b}}_d \end{pmatrix}$$

with $D = \mathrm{diag}(\lambda_1/\|\widehat{\mathbf{v}}_1\|, \ldots, \lambda_m/\|\widehat{\mathbf{v}}_d\|)$, and $\|\frac{\partial \rho(B)}{\partial \mathbf{t}}\|^2 = \sum_{i=1}^m \sum_{a=1}^d \lambda_i^2 \widehat{B}_{ia}^2 / \|\widehat{\mathbf{v}}_i\|^2 = \|\lambda\|^2$. Taking inner product between Eq. (18) and $\mathbf{u}_{ij}$ derives

$$\begin{aligned} \gamma_{ij} &= (2 - \delta_{ij}) \sum_{k=1}^m \lambda_k \frac{\widehat{B}_{ki}\widehat{B}_{kj}}{\|\tilde{\mathbf{v}}_k\|} + O_p(n^{-\tau}) \\ &= (2 - \delta_{ij}) \sum_{k=1}^q \lambda_k \frac{\widehat{B}_{ki}\widehat{B}_{kj}}{\|\tilde{\mathbf{v}}_k\|} + (2 - \delta_{ij}) \sum_{k=q+1}^m \lambda_k \widehat{B}_{ki} \frac{\widehat{B}_{kj}}{\|\tilde{\mathbf{v}}_k\|} + O_p(n^{-\tau}). \end{aligned}$$

From Theorem 3, the first term of the last line is of $O_p(\alpha_n)$ and the second term is of $O_p(\|\lambda\| n^{-\tau})$. The norm of the left hand side of Eq. (18) is $\|\lambda\|$, while the norm of the right hand side is $O_p(\alpha_n + \|\lambda\| n^{-\tau} + n^{-\tau})$. Since $\|\lambda\| \ge \beta_n$, it contradicts with the assumption $\alpha_n \ll n^{-\tau} \ll \beta_n$. There exists, therefore, $j \ge q + 1$ such that $\widehat{\mathbf{v}}_j = 0$ with probability tending to one.

The rest of the proof is done in exactly the same deduction argument as the proof of Theorem 2 in Chen et al. [3], and we omit it. □

E Detailed description of Boston Housing data

	Variable	Description
y	MEDV	Median value of homes in thousands of dollars
x^1	CRIM	Crime rate
x^2	ZN	Proportion of residential land zoned for lots over 25,000 sq. ft.
x^3	INDUS	Proportion of non-retail business acres (proxy for industry)
x^4	CHAS	Dummy variable indicating proximity to Charles River
x^5	NOX	Nitrogen oxide concentrations
x^6	RM	Average number of rooms
x^7	AGE	Proportion of units built prior to 1940
x^8	DIS	Weighted distances to major employment centers in area
x^9	RAD	Index of accessibility to radial highways
x^{10}	TAX	Property tax rate
x^{11}	PTRATIO	Pupil-Teacher ratio
x^{12}	B	Black proportion of population
x^{13}	LSTAT	Proportion of population that is lower socioeconomic status

References

[1] C.R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973.

[2] A. Caponnetto and E. De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[3] X. Chen, C. Zou, and R. Dennis Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Stat.*, 38(6):3696–3723, 2010.

[4] F. Bauer, S. Pereverzev and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

[5] K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *JMLR*, 5:73–99, 2004.

[6] J.H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Trans. Signal Processing*, 50(3):635–650, 2002.

[7] T. Rapcsák. On minimization on stiefel manifolds. *Euro. J. Operational Research*, 143(2):365 – 376, 2002.

[8] S. Smale, D.. Zhou. Shannon sampling II: Connections to learning theory. Applied and Computational Harmonic Analysis, Vol. 19, No. 3. (November 2005), pp. 285-302.

[9] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Appxoximation*, 26:153–172, 2007.