# HOW ARE THE CENTERED KERNEL PRINCIPAL COMPONENTS RELEVANT TO REGRESSION TASK? — AN EXACT ANALYSIS

*Masahiro Yukawa*[1,2*] *Klaus-Robert Müller*[3,4,5*] *Yuto Ogino*[1]

1. Department of Electronics and Electrical Engineering, Keio University, Japan
2. Center for Advanced Intelligence Project, RIKEN, Japan
3. Department of Machine Learning, Berlin Institute of Technology, Germany
4. Department of Brain and Cognitive Engineering, Korea University, Republic of Korea
5. Max Planck Institute for Informatics, Germany

## ABSTRACT

We present an exact analytic expression of the contributions of the kernel principal components to the relevant information in a nonlinear regression problem. A related study has been presented by Braun, Buhmann, and Müller in 2008, where an upper bound of the contributions was given for a general supervised learning problem but with "uncentered" kernel PCAs. Our analysis clarifies that the relevant information of a kernel regression under explicit centering operation is contained in a finite number of leading kernel principal components, as in the "uncentered" kernel-PCA case, if the kernel matches the underlying nonlinear function so that the eigenvalues of the centered kernel matrix decay quickly. We compare the regression performances of the least-square-based methods with the centered and uncentered kernel PCAs by simulations.

***Index Terms—*** nonlinear regression, kernel PCA, reproducing kernel Hilbert space, spectral decomposition

## 1. INTRODUCTION

Kernel principal component analysis (PCA) [1] is a nonlinear technique to find the principal components in a so-called *feature space* to which the sample vectors are mapped "nonlinearly" with a reproducing kernel. It can be used for dimensionality reduction, noise reduction, pre-processing of classification/regression/clustering, etc. (see, e.g., [2]). Although kernel PCA only looks at the samples, it is actually relevant to a task (a supervised learning problem such as classification/regression). This has been shown theoretically by Braun, Buhmann, and Müller [3]. More specifically, it has been shown that the *relevant information* about the task is typically contained in the subspace generated by a small number of leading kernel PCA components. Here, the relevant information is a noise-free version of the output (label) vector. The derivation therein is rather complicated since the integral operator associated with the reproducing kernel is considered first and its spectral decomposition is then truncated to attain a bound. This makes the analysis applicable only to an "uncentered" kernel PCA. We break this limitation in this paper. Uncentered (linear) PCAs have been used in different fields such as climatology [4], neuroimaging data [5], microarrays [6], among many others. Its use in reduced-rank signal processing has been studied, e.g., by Scharf [7]. Cadima and Jolliffe have studied the relationships between uncentered and centered (i.e., standard) PCAs, in particular the relationships between the eigenvalues of the two PCA variants [8]. However, the relationships between uncentered and centered (linear/kernel) PCAs have not been well studied in the context of regression.

In this paper, we focus on a nonlinear regression problem, and study how much the eigenfunctions of the covariance operator of kernel feature vectors contribute to the relevant information. Here, the contribution of an eigenfunction is the quantity of which an upper bound has been given in [3, 9] for an uncentered kernel PCA. We present its exact analytic expression (instead of its bound) for both centered and uncentered kernel PCAs. The contribution essentially decays at the same rate as the eigenvalues. A numerical example shows that the centered and uncentered kernel PCAs have a similar tendency in test-error-decay characteristics.

*This work was supported in part by JSPS Grants-in-Aids (15K06081, 15K13986, 15H02757), in part by the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology in the BK21 program, in part by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451), and in part by the German Research Foundation under Grant DFG MU 987/6-1, Grant SPP 1527, and Grant MU 987/14-1 as well as BMBF (BBDC).

## 2. CONTRIBUTIONS OF PRINCIPAL EIGENVECTORS IN LINEAR REGRESSION

In this section, we discuss the linear PCA of which the results are useful to discuss kernel PCA. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be a set of $n$ sample vectors $\boldsymbol{x}_i \in \mathbb{R}^N$. We consider a simple linear regression model

$$y_i := \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{w} + \epsilon_i,\ i = 1, 2, \cdots, n, \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^N$ is the regression vector, $\epsilon_i$ is the zero-mean additive noise, and $y_i$ is the output corresponding to the sample vector $\boldsymbol{x}_i$ ($(\cdot)^{\mathsf{T}}$ stands for transposition). The regression model in (1) can be written in a vector form as

$$\boldsymbol{y} := [y_1, y_2, \cdots, y_n]^{\mathsf{T}} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{X} := [\boldsymbol{x}_1\ \boldsymbol{x}_2 \cdots \boldsymbol{x}_n] \in \mathbb{R}^{N\times n}$ and $\boldsymbol{\epsilon} := [\epsilon_1, \epsilon_2, \cdots, \epsilon_n]^{\mathsf{T}}$. Define *the relevant information vector* [3]

$$\boldsymbol{g} := [g_1, g_2, \cdots, g_n] := [E(y_1|\boldsymbol{x}_1), E(y_2|\boldsymbol{x}_2), \cdots, E(y_n|\boldsymbol{x}_n)]^{\mathsf{T}}, \tag{3}$$

which is a noise-free version of $\boldsymbol{y}$. In the present regression case, it holds that

$$\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{\epsilon}. \tag{4}$$

Given any closed linear subspace $\mathcal{S}$ of an arbitrary real Hilbert space $\mathcal{X}$, we denote by $P_{\mathcal{S}}$ the orthogonal projection operator onto $\mathcal{S}$ that maps any point $x \in \mathcal{X}$ to its closest point in $\mathcal{S}$.

### 2.1. Uncentered linear PCA case

An uncentered linear PCA is carried out through the singular value decomposition of the uncentered sample matrix, which is denoted as

$$\boldsymbol{X} := [\boldsymbol{x}_1\ \boldsymbol{x}_2 \cdots \boldsymbol{x}_n] = \boldsymbol{U}_{\mathrm{u}}\boldsymbol{\Sigma}_{\mathrm{u}}\boldsymbol{V}_{\mathrm{u}}^{\mathsf{T}}, \tag{5}$$

where $\boldsymbol{U}_{\rm u} := [\boldsymbol{u}_1^{[\rm u]}\ \boldsymbol{u}_2^{[\rm u]} \cdots \boldsymbol{u}_N^{[\rm u]}] \in \mathbb{R}^{N\times N}$ and $\boldsymbol{V}_{\rm u} := [\boldsymbol{v}_1^{[\rm u]}\ \boldsymbol{v}_2^{[\rm u]} \cdots \boldsymbol{v}_n^{[\rm u]}] \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and, if $n > N$ for instance, $\boldsymbol{\Sigma}_{\rm u} := [\mathrm{diag}(\sigma_1^{[\rm u]}, \sigma_2^{[\rm u]}, \cdots, \sigma_N^{[\rm u]})\ \ \boldsymbol{O}_{N\times(n-N)}] \in \mathbb{R}^{N\times n}$. Here, $\boldsymbol{O}_{N\times(n-N)}$ is the $N \times (n-N)$ zero matrix.

We express $\boldsymbol{w}$ as

$$\boldsymbol{w} = \sum_{i=1}^{N} \beta_i \boldsymbol{u}_i^{[\rm u]} = \boldsymbol{U}_{\rm u}\boldsymbol{\beta}, \tag{6}$$

where $\boldsymbol{\beta} := [\beta_1, \beta_2, \cdots, \beta_N]^{\mathsf{T}} \in \mathbb{R}^N$. Plugging (5) and (6) into (2), we obtain

$$\boldsymbol{g} = \boldsymbol{V}_{\rm u}\boldsymbol{\Sigma}_{\rm u}\boldsymbol{U}_{\rm u}^{\mathsf{T}}\boldsymbol{U}_{\rm u}\boldsymbol{\beta} = \sum_{i=1}^{n} \beta_i \sigma_i^{[\rm u]} \boldsymbol{v}_i^{[\rm u]}. \tag{7}$$

Define the principal subspace $\mathcal{M}_d^{[\rm u]} := \mathrm{span}\{\boldsymbol{u}_1^{[\rm u]}, \boldsymbol{u}_2^{[\rm u]}, \cdots, \boldsymbol{u}_d^{[\rm u]}\} \subset \mathbb{R}^N$. For $\boldsymbol{X}_d := \left[P_{\mathcal{M}_d^{[\rm u]}}(\boldsymbol{x}_1), P_{\mathcal{M}_d^{[\rm u]}}(\boldsymbol{x}_2), \cdots, P_{\mathcal{M}_d^{[\rm u]}}(\boldsymbol{x}_n)\right] \in \mathbb{R}^{N\times d}$, it holds [10] that $\boldsymbol{X}_d^{\mathsf{T}}\hat{\boldsymbol{w}} = \boldsymbol{X}^{\mathsf{T}} P_{\mathcal{M}_d^{[\rm u]}}(\hat{\boldsymbol{w}})$, from which the following equivalence can be verified:

$$\min_{\hat{\boldsymbol{w}}\in\mathbb{R}^N} (\boldsymbol{y}-\boldsymbol{X}_d^{\mathsf{T}}\hat{\boldsymbol{w}})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X}_d^{\mathsf{T}}\hat{\boldsymbol{w}}) \Leftrightarrow \min_{\hat{\boldsymbol{w}}\in\mathcal{M}_d^{[\rm u]}} (\boldsymbol{y}-\boldsymbol{X}^{\mathsf{T}}\hat{\boldsymbol{w}})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{X}^{\mathsf{T}}\hat{\boldsymbol{w}}).$$

A simple inspection of (7) implies that $|\beta_i|\,\sigma_i^{[\rm u]} (\geq 0)$ expresses the amount of 'information' that is lost by excluding $\boldsymbol{u}_i^{[\rm u]}$ in the estimation of $\boldsymbol{w}$. Since $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{u}_i^{[\rm u]} = \beta_i$ from (6), one can verify with (7) that

$$|\boldsymbol{g}^{\mathsf{T}}\boldsymbol{v}_i^{[\rm u]}| = |\beta_i|\,\sigma_i^{[\rm u]} = \sigma_i^{[\rm u]}|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{u}_i^{[\rm u]}|. \tag{8}$$

### 2.2. Centered linear PCA case

Define the centered sample vectors $\boldsymbol{\xi}_i := \boldsymbol{x}_i - \bar{\boldsymbol{x}} \in \mathbb{R}^N$, $i = 1, 2, \cdots, n$, where $\bar{\boldsymbol{x}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$, and let $\boldsymbol{\Xi} := [\boldsymbol{\xi}_1\ \boldsymbol{\xi}_2 \cdots \boldsymbol{\xi}_n]$. Centered linear PCA is carried out through the singular value decomposition of $\boldsymbol{\Xi}$, which is denoted as

$$\boldsymbol{\Xi} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}, \tag{9}$$

where $\boldsymbol{U} := [\boldsymbol{u}_1\ \boldsymbol{u}_2 \cdots \boldsymbol{u}_N] \in \mathbb{R}^{N\times N}$ and $\boldsymbol{V} := [\boldsymbol{v}_1\ \boldsymbol{v}_2 \cdots \boldsymbol{v}_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and, if $n > N$ for instance, $\boldsymbol{\Sigma} := \left[\mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_N)\ \ \boldsymbol{O}_{N\times(n-N)}\right] \in \mathbb{R}^{N\times n}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$.

The mean output $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i = \bar{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{w} + \bar{\epsilon}$, where $\bar{\epsilon} := \frac{1}{n}\sum_{i=1}^n \epsilon_i$, contains the component originated from the mean-input vector $\bar{\boldsymbol{x}}$ (on top of the noise mean). As this component is common to all the outputs, we subtract it from $y_i$ as

$$\tilde{y}_i := y_i - \bar{y} = \boldsymbol{\xi}_i^{\mathsf{T}}\boldsymbol{w} + \tilde{\epsilon}_i, \quad i = 1, 2, \cdots, n, \tag{10}$$

where $\tilde{\epsilon}_i := \epsilon_i - \bar{\epsilon}$. Note that, once an estimate $\hat{\tilde{y}}_i$ of $\tilde{y}_i$ becomes available, $y_i$ can easily be estimated as $\hat{y}_i := \hat{\tilde{y}}_i + \bar{y}$. The 'centered' output $\tilde{y}_i$ is originated from the 'deviation' vector $\boldsymbol{\xi}_i$, excluding the noise deviation $\tilde{\epsilon}_i$. One may thus achieve efficient dimensionality-reduction, without losing essential information for estimating $\tilde{y}_i$, by projecting $\boldsymbol{\xi}_i$ onto a principal subspace of $\boldsymbol{\Xi}$.

We now define the *essential relevant information vector* $\tilde{\boldsymbol{g}} := [\tilde{g}_1, \tilde{g}_2, \cdots, \tilde{g}_n]^{\mathsf{T}}$ with

$$\tilde{g}_i := E(\tilde{y}_i|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) = \boldsymbol{\xi}_i^{\mathsf{T}}\boldsymbol{w}, \quad i = 1, 2, \cdots, n. \tag{11}$$

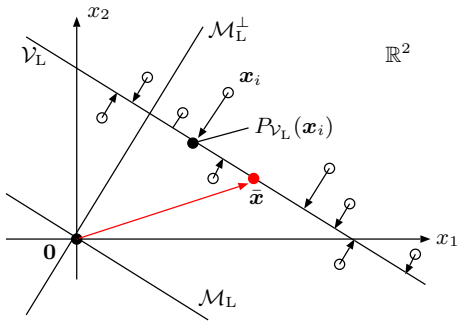



**Fig. 1**. An illustration of centered PCA.

Note here that $E(\tilde{\epsilon}_i) = 0$ due to the zero-mean noise assumption. Since $\tilde{\boldsymbol{g}} = \boldsymbol{\Xi}^{\mathsf{T}}\boldsymbol{w}$, we obtain

$$|\tilde{\boldsymbol{g}}^{\mathsf{T}}\boldsymbol{v}_i| = |\sigma_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{u}_i)|. \tag{12}$$

Centered PCA is illustrated in Fig. 1, where $\mathcal{V}_{\rm L} := \mathcal{M}_{\rm L} + \bar{\boldsymbol{x}} := \{\boldsymbol{x} + \bar{\boldsymbol{x}} \mid \boldsymbol{x} \in \mathcal{M}_{\rm L}\}$ with the principal subspace $\mathcal{M}_{\rm L} := \mathrm{span}\{\boldsymbol{u}_i\}_{i=1}^d$ of dimension $d$. In the figure, $\mathcal{M}_{\rm L}^{\perp}$ is the orthogonal complement of the subspace $\mathcal{M}_{\rm L}$. From (12), it is seen that the essential contribution decays at the same rate as $\sigma_i$, as in the uncentered case.

## 3. CONTRIBUTIONS OF PRINCIPAL EIGENFUNCTIONS IN NONLINEAR REGRESSION

Let $\kappa : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be a reproducing (or equivalently positive definite) kernel. A typical example of reproducing kernel is the Gaussian kernel: $\kappa(\boldsymbol{x}, \boldsymbol{y}) := \exp[-(\boldsymbol{x}-\boldsymbol{y})^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{y})/2\sigma^2]$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, for some positive constant $\sigma > 0$. We denote by $\mathcal{H}$ the reproducing kernel Hilbert space (RKHS) associated with $\kappa$. The nonlinear mapping $\Phi : \mathbb{R}^N \to \mathcal{H}$, $\boldsymbol{x} \mapsto \phi_{\boldsymbol{x}}$ is often called *the feature map*, where $\phi_{\boldsymbol{x}}(\boldsymbol{y}) = \kappa(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}$, $\boldsymbol{y} \in \mathbb{R}^N$. Denote by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ the inner product and the norm defined in $\mathcal{H}$, respectively.

### 3.1. Nonlinear regression model and linearization

We consider a nonlinear regression model

$$y_i := \psi(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{13}$$

We assume here that $\psi \in \mathcal{H}$ so that $\psi(\boldsymbol{x}_i) = \langle\psi, \Phi(\boldsymbol{x}_i)\rangle$. Thus, $\Phi(\boldsymbol{x}_i)$ is an "uncentered" feature vector (or an "uncentered" sample vector in the feature space $\mathcal{H}$). Define the closed subspace $\mathcal{M} := \mathrm{span}\{\Phi(\boldsymbol{x}_i)\}_{i=1}^n$, and denote its orthogonal complement by $\mathcal{M}^{\perp} := \{f \in \mathcal{H} \mid \langle f, g\rangle = 0,\ \forall g \in \mathcal{M}\}$. Under the orthogonal decomposition [10] $\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^{\perp}$, the vector $\psi$ can be decomposed as

$$\psi = P_{\mathcal{M}}(\psi) + P_{\mathcal{M}^{\perp}}(\psi). \tag{14}$$

As $\Phi(\boldsymbol{x}_i) \in \mathcal{M}$ for all $i = 1, 2, \cdots, n$ by definition of $\mathcal{M}$, it follows that

$$\psi(\boldsymbol{x}_i) = \langle P_{\mathcal{M}}(\psi), \Phi(\boldsymbol{x}_i)\rangle\,. \tag{15}$$

The vector $P_{\mathcal{M}}(\psi)$ on $\mathcal{M}$ can be expressed as

$$P_{\mathcal{M}}(\psi) = \sum_{j=1}^{n} \alpha_j \Phi(\boldsymbol{x}_j) \tag{16}$$

for the coefficients $\alpha_j, j = 1, 2, \cdots, n$. By (15) and (16), it follows that

$$\psi(\boldsymbol{x}_i) = \sum_{j=1}^{n} \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{k}_i^{[\mathrm{u}]}, \tag{17}$$

where $\boldsymbol{k}_i^{[\mathrm{u}]} := [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_1), \kappa(\boldsymbol{x}_i, \boldsymbol{x}_2), \cdots, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_n)]^\mathsf{T}$ and $\boldsymbol{\alpha} := [\alpha_1, \alpha_2, \cdots, \alpha_n]^\mathsf{T}$. By (13) and (17), the output vector can be written as

$$\boldsymbol{y} = \boldsymbol{K}_\mathrm{u} \boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{18}$$

where the kernel matrix $\boldsymbol{K}_\mathrm{u} := [\boldsymbol{k}_1^{[\mathrm{u}]}\ \boldsymbol{k}_2^{[\mathrm{u}]} \cdots \boldsymbol{k}_n^{[\mathrm{u}]}]$ is symmetric and positive semi-definite.

### 3.2. Uncentered kernel PCA case

We consider an uncentered kernel PCA based on the eigenvalue decomposition of the kernel matrix:

$$\boldsymbol{K}_\mathrm{u} = \boldsymbol{Q}_\mathrm{u} \boldsymbol{\Lambda}_\mathrm{u} \boldsymbol{Q}_\mathrm{u}^\mathsf{T}, \tag{19}$$

where $\boldsymbol{Q}_\mathrm{u} := [\boldsymbol{q}_1^{[\mathrm{u}]}\ \boldsymbol{q}_2^{[\mathrm{u}]} \cdots \boldsymbol{q}_n^{[\mathrm{u}]}] \in \mathbb{R}^{n \times n}$ is an orthogonal matrix and $\boldsymbol{\Lambda}_\mathrm{u} := \mathrm{diag}(\lambda_1^{[\mathrm{u}]}, \lambda_2^{[\mathrm{u}]}, \cdots, \lambda_n^{[\mathrm{u}]}) \in \mathbb{R}^{n \times n}$ with $\lambda_1^{[\mathrm{u}]} \geq \lambda_2^{[\mathrm{u}]} \geq \cdots \geq \lambda_n^{[\mathrm{u}]} \geq 0$. Comparing (18) with (2) under (19), we immediately obtain from (8) the following relation:

$$|\boldsymbol{g}^\mathsf{T} \boldsymbol{q}_i^{[\mathrm{u}]}| = \lambda_i^{[\mathrm{u}]} |\boldsymbol{\alpha}^\mathsf{T} \boldsymbol{q}_i^{[\mathrm{u}]}|. \tag{20}$$

Here, $\boldsymbol{g} = [\psi(\boldsymbol{x}_1), \psi(\boldsymbol{x}_2), \cdots, \psi(\boldsymbol{x}_n)]^\mathsf{T}$ in this case (see (3) for its definition).

### 3.3. Centered kernel PCA case

We consider the feature vector centered in the RKHS $\mathcal{H}$:

$$\phi_i := \Phi(\boldsymbol{x}_i) - \bar{\phi}, \quad i = 1, 2, \cdots, n, \tag{21}$$

where $\bar{\phi} := \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i)$. It then holds that

$$P_{\mathcal{M}}(\psi) = \eta \bar{\phi} + \sum_{j=1}^{n} \alpha_j \phi_j, \tag{22}$$

where $\eta := \sum_{j=1}^{n} \alpha_j$. Note that $\bar{\phi} \notin \mathcal{M}_\phi := \mathrm{span}\{\phi_i\}_{i=1}^n$ in general (see Appendix). Kernel PCA is based on the spectral decomposition of the sample covariance operator[1]

$$\boldsymbol{C} := \sum_{i=1}^{n} \phi_i \otimes \bar{\phi}_i, \tag{23}$$

where $\otimes$ denotes the Schatten product,[2] i.e., $(\phi_i \otimes \bar{\phi}_i) f := \langle \phi_i, f \rangle \phi_i$ for any $f \in \mathcal{H}$. Since $\boldsymbol{C}$ is a finite-dimensional operator (i.e., its image has a finite dimension at most $n$), it is a compact operator (or a completely continuous operator).[3] Hence, the spectral representation theorem admits the following representation:

$$\boldsymbol{C} = \sum_{i=1}^{r} \lambda_i p_i \otimes \bar{p}_i, \tag{24}$$

where $\{p_i\}_{i=1}^r \subset \mathcal{H}$ is an orthonormal basis of the $r$-dimensional subspace $\mathcal{M}_\phi$, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ (see Appendix).

It is known that the eigenvalue problem of $\boldsymbol{C}$ is a dual problem of the eigenvalue problem of $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ with its $(i, j)$ entry $k_{i,j} := \langle \phi_i, \phi_j \rangle$. The eigenvalue decomposition of $\boldsymbol{K}$ is denoted as

$$\boldsymbol{K} =: [\boldsymbol{k}_1\ \boldsymbol{k}_2 \cdots \boldsymbol{k}_n]^\mathsf{T} = \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^\mathsf{T}, \tag{25}$$

where $\boldsymbol{Q} := [\boldsymbol{q}_1\ \boldsymbol{q}_2 \cdots \boldsymbol{q}_n] \in \mathbb{R}^{n \times n}$ is an orthogonal matrix, and $\boldsymbol{\Lambda} := \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \in \mathbb{R}^{n \times n}$, where $\lambda_i := 0$ for $i = r + 1, \cdots, n$. Define a bounded linear operator

$$A : \mathcal{H} \to \mathbb{R}^n,\ f \mapsto [\langle \phi_1, f \rangle, \langle \phi_2, f \rangle, \cdots, \langle \phi_n, f \rangle]^\mathsf{T}. \tag{26}$$

Its spectral representation is then given by

$$A = \sum_{i=1}^{r} \sqrt{\lambda_i} \boldsymbol{q}_i \otimes \bar{p}_i. \tag{27}$$

Denoting by $A^*$ the adjoint operator of $A$, one can verify that $\boldsymbol{C} = A^* \circ A$ and $\boldsymbol{K} = A \circ A^*$, where the latter equation holds when one regards $\boldsymbol{K}$ as a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^n$ (the symbol $\circ$ denotes composition of operators). Since $\langle Af, \boldsymbol{x} \rangle = \sum_{i=1}^n x_i \langle \phi_i, f \rangle = \langle f, \sum_{i=1}^n x_i \phi_i \rangle$ for any $f \in \mathcal{H}$ and $\boldsymbol{x} := [x_1, x_2, \cdots, x_N]^\mathsf{T} \in \mathbb{R}^N$, it holds that $A^* \boldsymbol{x} = \sum_{i=1}^n x_i \phi_i$. It then follows immediately that $\boldsymbol{K}\boldsymbol{1} = A(A^* \boldsymbol{1}) = A(\sum_{i=1}^n \phi_i) = A(0) = \boldsymbol{0}$, where $\boldsymbol{1} := [1, 1, \cdots, 1]^\mathsf{T} \in \mathbb{R}^n$. This implies that the matrix $\boldsymbol{K}$ is rand-deficient, and thus $r \leq n - 1$ (cf. Appendix).[4]

In the present nonlinear case, the centered output defined in (10) is given by

$$\tilde{y}_i = \langle \psi, \phi_i \rangle + \tilde{\epsilon}_i, \quad i = 1, 2, \cdots, n. \tag{28}$$

Accordingly, each component of $\tilde{\boldsymbol{g}}$ (see (11)) is given by

$$\tilde{g}_i = \langle \psi, \phi_i \rangle = \left\langle P_{\mathcal{M}_\phi}(\psi), \phi_i \right\rangle, \quad i = 1, 2, \cdots, n, \tag{29}$$

where the second equality is verified with $\phi_i \in \mathcal{M}_\phi$. We are now ready to show the following theorem.

**Theorem 1** *Let $P_{\mathcal{M}_\phi}(\psi) = \sum_{i=1}^n \gamma_i \phi_i = A^* \boldsymbol{\gamma} \in \mathcal{M}_\phi$, where $\boldsymbol{\gamma} := [\gamma_1, \gamma_2, \cdots, \gamma_n]^\mathsf{T} \in \mathbb{R}^n$. Then, the contribution of $\boldsymbol{q}_i$ to the essential relevant information vector $\tilde{\boldsymbol{g}}$ is given by*

$$|\tilde{\boldsymbol{g}}^\mathsf{T} \boldsymbol{q}_i| = \lambda_i |\boldsymbol{\gamma}^\mathsf{T} \boldsymbol{q}_i|. \tag{30}$$

Proof. Substituting $P_{\mathcal{M}_\phi}(\psi) = A^* \boldsymbol{\gamma}$ into (29), we obtain $\tilde{g}_i = \langle A^* \boldsymbol{\gamma}, \phi_i \rangle$ and thus

$$\tilde{\boldsymbol{g}} = A \circ A^*(\boldsymbol{\gamma}) = \boldsymbol{K} \boldsymbol{\gamma}. \tag{31}$$

By (25) and (31), one can verify the assertion. □

Theorem 1 states that the contribution of $\boldsymbol{q}_i$ to the essential relevant information vector $\tilde{\boldsymbol{g}}$ decays at the same rate as $\lambda_i$, as in the case of the uncentered kernel PCA.

[1] The sample covariance operator is also expressed in the following way [2]: $\boldsymbol{C} := \frac{1}{n} \sum_{i=1}^n (\Phi(\boldsymbol{x}_i) - \bar{\phi})(\Phi(\boldsymbol{x}_i) - \bar{\phi})^\mathsf{T}$.

[2] The bar on $\phi_i$ is a part of the Schatten product.

[3] A linear operator is said to be a compact operator (or a completely continuous operator) if it maps a bounded set to a compact set.

[4] Let $p = \sum_{i=1}^n q_i \phi_i \in \mathcal{H}$, $q_i \in \mathbb{R}$, $i = 1, 2, \cdots, n$. Then, $\lambda p = \boldsymbol{C} p$, $\lambda \in \mathbb{R}$, if and only if $\lambda \boldsymbol{K} \boldsymbol{q} = \boldsymbol{K}^2 \boldsymbol{q}$, where $\boldsymbol{q} := [q_1, q_2, \cdots, q_n]^\mathsf{T}$. Therefore, any solution $\boldsymbol{q}$ of the dual eigenvalue problem $\lambda \boldsymbol{q} = \boldsymbol{K} \boldsymbol{q}$ gives $p$ such that $\lambda p = \boldsymbol{C} p$ (see [2]). However, $\boldsymbol{q}$ corresponding to $\lambda = 0$ may give $p = 0$, as is the case of $\boldsymbol{q} = \boldsymbol{1}$. In such a case, the solution $\boldsymbol{q}$ does not give a solution to the primal eigenvalue problem, since $p = 0$ is not an eigenvector by definition.

### 3.4. Discussion

**Relation to prior work:** We discuss the relation between (20) and the result in [3, Theorem 1]. Referring to (20), one can see that the right hand side contains a sole term which decays at the same rate as the eigenvalue $\lambda_i$. The truncation error contained in the result in [3] does not appear in our analysis due to the simple derivation without using the integral operator. The absence of the truncation error results in the exact equality appearing in (20) (unlike the result in [3] which shows an upper bound of $|\boldsymbol{g}^\mathsf{T}\boldsymbol{q}_i|$ essentially). One may concern a situation when the values of $|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{q}_i|$ are small for the leading eigenvectors compared to those for minor eigenvectors. This situation however corresponds to the case where many of the sample vectors $\boldsymbol{x}_i$ are filtered out by the nonlinear function $\psi$ and accordingly many of the outputs $y_i$ contain little signals. In such a situation, the learning problem itself is very challenging, and it would be advised to attain some extra sample vectors that are not filtered out by $\psi$. It should be mentioned that the decay rate of $\lambda_i$ depends on the choice of the kernel (see [3]).

## 4. NUMERICAL EXAMPLE

We conduct a computer simulation using toy data of dimension $N = 2$ to illustrate performance of the variants of kernel PCA: the uncentered kernel PCA and the centered kernel PCA. Training dataset is generated as follows:

1. specify five different cluster centers $\boldsymbol{c}_1 := (0, 0)$, $\boldsymbol{c}_2 := (5, 4)$, $\boldsymbol{c}_3 := (-3, 2)$, $\boldsymbol{c}_4 := (1, 4)$, $\boldsymbol{c}_5 := (10, -5)$, and
2. generate 100 data points randomly around one of the cluster centers from the i. i. d. normal distribution $\mathcal{N}(\boldsymbol{c}_i, 0.04\boldsymbol{I})$, where the cluster center is chosen randomly with equal probability for each data point.

Test dataset is generated in exactly the same way as (but independently from) the training data. The noise $\epsilon$ obeys the i.i.d. normal distribution $\mathcal{N}(0, 0.01)$. The nonlinear function $\psi$ and the training data are depicted in Figure 2. In kernel PCA, the Gaussian kernel is used for $\sigma := 0.2$.

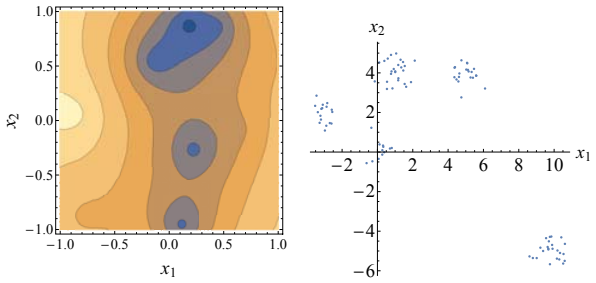
Figure 3 plots (a) the singular values of the kernel matrix $\boldsymbol{K}_\mathrm{u}$ (Uncentered) and $\boldsymbol{K}$ (Centered), and (b) the estimation error $\sqrt{(\boldsymbol{y}_\mathrm{test} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y}_\mathrm{test} - \hat{\boldsymbol{y}})/\boldsymbol{y}_\mathrm{test}^\mathsf{T}\boldsymbol{y}_\mathrm{test}}$, where $\boldsymbol{y}_\mathrm{test}$ is the ouput vector for the test data and $\hat{\boldsymbol{y}}$ is its estimate. One can see that there is no notable difference in the singular-value decay between the two cases. Referring to Figure 3(b), the centered and uncentered kernel PCAs show a similar tendency.

## 5. CONCLUSION

We studied the contributions of the kernel principal components to the relevant information in nonlinear regression for both centered and uncentered cases. Our simple derivation enabled to deal with the centering operation explicitly and led to the exact analytic expression of the contribution (containing no truncation errors). It turned out that the essential contribution decayed at the same rate as the eigenvalues. The numerical example showed that the centered and uncentered kernel PCAs had a similar tendency in test-error-decay characteristics. The present study will be useful to estimate the relevant dimension in nonlinear regression under the centering operation, as well as improving the performance of kernel adaptive filtering [11–17].

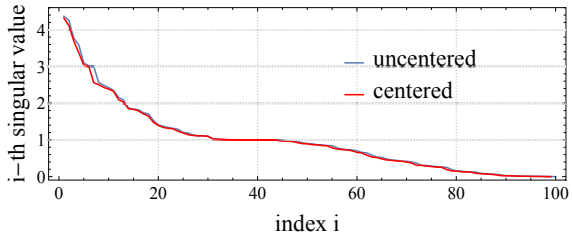
## APPENDIX

By definition of $\boldsymbol{C}$, the number $r$ of nonzero eigenvalues of $\boldsymbol{C}$ equals to the dimension of $\mathcal{M}_\phi (:= \mathrm{span}\{\phi_i\}_{i=1}^n)$. We can verify the following lemmas. (All the lemmas shown below apply to linear PCA. We omit the proof of the second lemma).

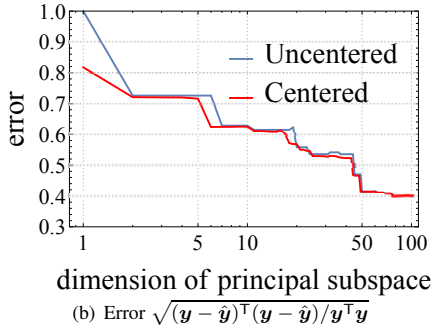



(a) Nonlinear function $\psi$ (b) Input data

**Fig. 2**. Simulation data.




(a) Singular-value decay




(b) Error $\sqrt{(\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})/\boldsymbol{y}^\mathsf{T}\boldsymbol{y}}$

**Fig. 3**. Simulation results.

**Lemma 1** *Assume that $\bar{\phi} \notin \mathcal{M}_\phi$. Then, the following statements hold.*

1. $\mathcal{M} = \mathcal{M}_\phi + \mathrm{span}\{\bar{\phi}\} := \{f + g \mid f \in \mathcal{M}_\phi, g \in \mathrm{span}\{\bar{\phi}\}\}$
2. $\dim(\mathcal{M}_\phi) = \dim(\mathcal{M}) - 1$.

Proof of 1: It is clear that $\Phi(\boldsymbol{x}_i) = \phi_i + \bar{\phi} \in \mathcal{M}_\phi + \mathrm{span}\{\bar{\phi}\}$, which implies that $\mathcal{M} \subset \mathcal{M}_\phi + \mathrm{span}\{\bar{\phi}\}$. On the other hand, it holds that $\phi_i, \bar{\phi} \in \mathcal{M}$, which implies that $\mathcal{M} \supset \mathcal{M}_\phi + \mathrm{span}\{\bar{\phi}\}$. This completes the proof.
Proof of 2: Clear from Lemma 1.1 and the assumption. □

**Lemma 2** *$\bar{\phi} \in \mathcal{M}_\phi$ if and only if $0 \in \mathrm{aff}\{\Phi(\boldsymbol{x}_i)\}_{i=1}^n$; i.e., there exist $\alpha_i \in \mathbb{R}$, $i = 1, 2, \cdots, n$, such that $\sum_{i=1}^n \alpha_i = 1$ and $\sum_{i=1}^n \alpha_i \Phi(\boldsymbol{x}_i) = 0$. Here, $\mathrm{aff}(\cdot)$ stands for the affine hull.*

**Lemma 3** *If $\{\Phi(\boldsymbol{x}_i)\}_{i=1}^n$ is linearly independent, then $\bar{\phi} \notin \mathcal{M}_\phi$.*

Proof. It is clear that $0 \in \mathrm{aff}\{\Phi(\boldsymbol{x}_i)\}_{i=1}^n \Leftrightarrow [\sum_{i=1}^n \alpha_i \Phi(\boldsymbol{x}_i) = 0$ for some $\alpha_i \in \mathbb{R}$ such that $\sum_{i=1}^n \alpha_i = 1] \Rightarrow [\{\Phi(\boldsymbol{x}_i)\}_{i=1}^n$ is linearly dependent]. Hence, the linear independence assumption implies $0 \notin \mathrm{aff}\{\Phi(\boldsymbol{x}_i)\}_{i=1}^n$ and thus $\bar{\phi} \notin \mathcal{M}_\phi$ by Lemma 2. □

Lemma 3 suggests that, when the Gaussian kernel is employed, $\bar{\phi} \notin \mathcal{M}_\phi$ and hence the statements of Lemma 1 hold true.

## 6. REFERENCES


[1] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[2] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[3] M. L. Braun, J. M. Buhmann, and K.-R. Müller, "On relevant dimensions in kernel feature spaces," *Journal of Machine Learning Research*, vol. 9, pp. 1875–1908, 2008.

[4] H. van den Dool, *Empirical methods in short-term climate prediction*, Oxford University Press, 2007.

[5] F. Gwadry, C. Berenstein, J. van Horn, and A. Braun, "Implementation and application of principal component analysis on functional neuroimaging data," Tech. Rep., Institute for Systems Research, 2001.

[6] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 2002.

[7] L. L. Scharf, "The SVD and reduced rank signal processing," *Signal Processing*, vol. 25, pp. 113–133, 1991.

[8] J. Cadima and I. Jolliffe, "On relationships between uncentred and column-centred principal component analysis," *Pakistan J. Statistics*, vol. 25, no. 4, pp. 473–503, 2009.

[9] G. Montavon, M. L. Braun, T. Krueger, and K.-R. Müller, "Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 62–74, July 2013.

[10] D. G. Luenberger, *Optimization by Vector Space Methods*, New York: Wiley, 1969.

[11] W. Liu, J. Príncipe, and S. Haykin, *Kernel Adaptive Filtering*, Wiley, New Jersey, 2010.

[12] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2165–2176, Aug. 2004.

[13] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2275–2285, Aug. 2004.

[14] C. Richard, J. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1058–1067, Mar. 2009.

[15] K. Slavakis, S. Theodoridis, and I. Yamada, "Adaptive constrained learning in reproducing kernel Hilbert spaces: the robust beamforming case," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4744–4764, Dec. 2009.

[16] M. Yukawa, "Multikernel adaptive filtering," *IEEE Trans. Signal Processing*, vol. 60, no. 9, pp. 4672–4682, Sept. 2012.

[17] M. Yukawa, "Adaptive learning in Cartesian product of reproducing kernel Hilbert spaces," *IEEE Trans. Signal Processing*, vol. 63, no. 22, pp. 6037–6048, Nov. 2015.