Mathematical formulation and application of kernel tensor decomposition based unsupervised feature extraction

Y-h. Taguchi[a,*], Turki Turki[b]

[a]Department of Physics, Chuo University, Tokyo 112-8551, Japan
[b]Department of Computer Science, King Abdulaziz University, Jeddah, 21589, Saudi Arabia

ARTICLE INFO

Keywords:
Kernel trick
Tensor decomposition
Unsupervised learning
Feature extraction
Applications in biology and medicine

ABSTRACT

In this work, we extended the recently developed tensor decomposition (TD) based unsupervised feature extraction (FE) to a kernel-based method through a mathematical formulation. Subsequently, the kernel TD (KTD) based unsupervised FE was applied to two synthetic examples as well as real data sets, and the findings were compared with those obtained previously using the TD-based unsupervised FE approaches. The KTD-based unsupervised FE outperformed or performed comparably with the TD-based unsupervised FE in *large p small n* situations, which are situations involving a limited number of samples with many variables (observations). Nevertheless, the KTD-based unsupervised FE outperformed the TD-based unsupervised FE in non *large p small n* situations. In general, although the use of the kernel trick can help the TD-based unsupervised FE gain more variations, a wider range of problems may also be encountered. Considering the outperformance or comparable performance of the KTD-based unsupervised FE compared to the TD-based unsupervised FE when applied to *large p small n* problems, it is expected that the KTD-based unsupervised FE can be applied in the genomic science domain, which involves many *large p small n* problems, and, in which, the TD-based unsupervised FE approach has been effectively applied.

Introduction

Recently, the tensor decomposition (TD) based and principal component analysis (PCA) based unsupervised feature extraction (FE) (Taguchi, 2020) approach was developed to identify a limited number of genes in *large p small n* problems involving a small number of samples (n) with a large number of features (genes) (p). This approach, applied to genomic science applications, which frequently involve *large p small n* problems, could successfully identify the limited number of biologically reliable genes that cannot be selected using conventional statistical test based feature selection methods. Nevertheless, because the TD and PCA based unsupervised FE do not include any tunable parameters owing to their linearity, the methods cannot be modified or optimized in failure scenarios. Thus, it is desirable to extend the TD and PCA based unsupervised FE to include nonlinearity.

To this end, we aimed at extending the TD-based unsupervised FE to incorporate the kernel trick (Schölkopf, 2000) and introduce nonlinearity. Because tensors do not have inner products that can be replaced with nonlinear kernels, we incorporate the self-inner products of tensors. In particular, the inner product is replaced with nonlinear kernels, and TD is applied to the generated tensor including nonlinear kernels. In this framework, the TD can be easily "kernelized."

Several researchers have attempted to apply kernel methods to process tensor data. For instance, Signoretto et al. (2011); Signoretto et al. (2012); Zhao et al. (2013a,b) input tensors into vectors (or matrices), which were then used to construct the kernels. However, such conversion may destroy the structural information of the tensor data. Moreover, the dimensionality of the resulting vector is typically extremely high, which leads to the curse of dimensionality and small sample size problems (Liu et al., 2015; Yan et al., 2007). He et al. (2017) proposed an implementation in which the tensor was reproduced using a kernelized TD.

It is expected that these problems can be avoided by computing the inner product of the tensors, as realized in the proposed approach.

*Corresponding author
tag@granular.com (Y-h. Taguchi)
ORCID(s): 0000-0003-0867-8986 (Y-h. Taguchi); 0000-0002-9491-2435 (T. Turki)

Results

Mathematical formulations

Suppose there exists a tensor $x_{ijk} \in \mathbb{R}^{N \times M \times K}$, which represents the value of the ith feature of the samples with properties j and k. For example, x_{ijk} might represent the price of product i that a person with a gender k and age j previously bought. In this case, N is the number of products available, M is the age, and $K \in [1, 2]$ represents male or female. Alternatively, in genomic science applications, which are the focus domain of this work, x_{ijk} is the expression of the ith genes of the kth tissue of the jth person; in this case, N is the number of genes, M is the number of participants in the study, and K is the number of tissues for which the expression of genes is to be measured. As another example, x_{ijk} might represent the electric current of the ith circuit at the jth temperature and kth atomic pressure; in this case, N is the number of circuits in a machine, and M and K denote the number of recordings of the temperature and atomic pressure, respectively.

The aim of TD-based unsupervised FE is to select the limited number of critical features among all the features (as many as N). In the first case, when i represents a product, the purpose of analysis may be to identify the limited number of products, the buyers of which are restricted to specific ages and genders. In the second case, when i represents a gene, the purpose of analysis may be to determine the limited number of genes whose altered expression may cause diseases. In the third example, when i is a circuit, the purpose of analysis may be to identify the limited number of circuits whose malfunctioning may result in the failure of the machine. Thus, in general, the purpose of TD-based unsupervised FE is to screen a small number of significant features from the larger set of features.

In this regard, in TD-based unsupervised FE, higher order singular value decomposition (Taguchi, 2020) (HOSVD) is applied to x_{ijk} to yield

$$x_{ijk} = \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \tag{1}$$

where $G(\ell_1 \ell_2 \ell_3) \in \mathbb{R}^{N \times M \times K}$ is a core tensor, which represents the weight of product $u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k}$ of the contribution to x_{ijk}. $u_{\ell_1 i} \in \mathbb{R}^{N \times N}, u_{\ell_2 j} \in \mathbb{R}^{M \times M}, u_{\ell_3 k} \in \mathbb{R}^{K \times K}$ denote singular-value vectors that represent the various dependencies of x_{ijk} on i, j, k, respectively; these vectors are orthogonal matrices.

To select the critical is, we must specify $u_{\ell_1 i}$ to select is; The is must be those for which the absolute values of $u_{\ell_1 i}$ are large. In this context, we must first identify $u_{\ell_2 j}$ and $u_{\ell_3 k}$, which represent the properties of interest. In the first example, the absolute value of $u_{\ell_2 j}$, which corresponds to the age, must be larger than a specific age. $u_{\ell_3 j}$, which corresponds to the genders, must be assigned distinct values for the two genders. In the second example, $u_{\ell_2 j}$, which corresponds to study participants, must be assigned distinct values for the patients and healthy controls. $u_{\ell_3 j}$, which corresponds to tissues, must have distinct values for the different tissues. In the third example, $u_{\ell_2 j}$ and $u_{\ell_3 k}$ must be assigned extremely large values at specific temperatures and atomic pressures.

After identifying the $u_{\ell_2 j}$ and $u_{\ell_3 k}$ of interest, we attempt to identify $G(\ell_1 \ell_2 \ell_3)$ that have larger absolute values given ℓ_2 and ℓ_3. Once ℓ_1 with G having the large absolute values is specified, we attempt to identify i with large absolute values of $u_{\ell_1 i}$. These is might represent the critical products bought by persons with specific ages and gender, genes expressive in certain specific tissues of patients, and circuits in which the electric current increases drastically at specific temperatures and atomic pressures.

The application of TD-based unsupervised FE to problems in genomic science yielded satisfactory results even when conventional feature selection methods based on statistical tests failed (Taguchi, 2020; Taguchi and Turki, 2020; Ng and Taguchi, 2020). Nevertheless, TD-based unsupervised FE has certain limitations. Specifically, owing to the tunable parameters, this approach cannot be modified or optimized in failure scenarios. The proposed method aims at extending the TD-based unsupervised FE to be incorporated with kernel tricks (Schölkopf, 2000). In this scenario, because many kernels can be selected, a better strategy applicable to the target problems can be identified.

To apply the kernel trick, TD must be suitably modified. First, we consider the partial sum as

$$x_{jkj'k'} = \sum_{i} x_{ijk} x_{ij'k'} \tag{2}$$

$x_{ij'k'}$ is equivalent to x_{ijk} with replacing j, k with j', k'. Substituting Eq. (1) into the aforementioned equation yields

$$\begin{aligned}
x_{jkj'k'} &= \sum_i \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \sum_{\ell'_1=1}^{N} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} G(\ell'_1\ell'_2\ell'_3) u_{\ell'_1 i} u_{\ell'_2 j'} u_{\ell'_3 k'} && (3)\\
&= \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} \sum_{\ell'_1=1}^{N} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} G(\ell_1\ell_2\ell_3) u_{\ell_2 j} u_{\ell_3 k} G(\ell'_1\ell'_2\ell'_3) u_{\ell'_2 j'} u_{\ell'_3 k'} \sum_i u_{\ell_1 i} u_{\ell'_1 i} && (4)\\
&= \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} \sum_{\ell'_1=1}^{N} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} G(\ell_1\ell_2\ell_3) u_{\ell_2 j} u_{\ell_3 k} G(\ell'_1\ell'_2\ell'_3) u_{\ell'_2 j'} u_{\ell'_3 k'} \delta_{\ell_1\ell'_1} && (5)\\
&= \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} \left(\sum_{\ell_1=1}^{N} G(\ell_1\ell_2\ell_3) G(\ell_1\ell'_2\ell'_3) \right) u_{\ell_2 j} u_{\ell_3 k} u_{\ell'_2 j'} u_{\ell'_3 k'} && (6)
\end{aligned}$$

Thus, we can obtain $u_{\ell_2 j}$ and $u_{\ell_3 k}$ by applying the HOSVD to Eq. (2) as

$$x_{jkj'k'} = \sum_{\ell_2} \sum_{\ell_3} \sum_{\ell'_2} \sum_{\ell'_3} G(\ell_2\ell_3\ell'_2\ell'_3) u_{\ell_2 j} u_{\ell_3 k} u_{\ell'_2 j} u_{\ell'_3 k}. \tag{7}$$

Note that when using the linear kernel, as indicated in Eq. (2), the KTD is equivalent to the (linear) TD.

Because Eq. (2) is expressed in the form of an inner product, it can be easily extended to (nonlinear) kernels. Although in the presented analysis we only considered the radial base function (RBF) kernel as an example, any other kernel can be used in place of the RBF kernel. Eq. (2) can be easily extended to the RBF kernel as

$$x_{jkj'k'} = \exp\left\{ -\alpha \sum_i \left(x_{ijk} - x_{ij'k'} \right)^2 \right\} \tag{8}$$

to which the HOSVD can be applied as is. This operation generates an expression corresponding to Eq. (7) with distinct Gs and $u_{\ell_2 j}, u_{\ell_3 k}, u_{\ell'_2 j}, u_{\ell'_3 k}$s.

To integrate two matrices x_{ij} and x_{kj} that share the sample j, we consider

$$x_{jj'} = \sum_{j''} \sum_i x_{ij} x_{ij''} \sum_k x_{kj''} x_{kj'} \tag{9}$$

to which the singular value decomposition (SVD) can be applied as is

$$x_{jj'} = \sum_\ell u_{\ell j} \lambda_\ell v_{\ell j'} \tag{10}$$

This expression can be easily extended to the RBF kernel as

$$x_{jj'} = \sum_{j''} \left[\exp\left\{ -\alpha \sum_i \left(x_{ij} - x_{ij''} \right)^2 \right\} \exp\left\{ -\alpha' \sum_k \left(x_{kj''} - x_{kj'} \right)^2 \right\} \right] \tag{11}$$

to which the SVD can be applied as is, and an expression similar to Eq. (10) can be obtained with distinct $u_{\ell j}, \lambda_\ell, v_{\ell j'}$

Finally, if two matrices, x_{ij} and x_{ik}, share feature i, we have

$$x_{jk} = \sum_i x_{ij} x_{ik} \tag{12}$$

to which the SVD can be applied as is, yielding

$$x_{jk} = \sum_\ell u_{\ell j} \lambda_\ell v_{\ell k}. \tag{13}$$

This expression can be easily extended to the RBF kernel as

$$x_{jk} = \exp\left\{-\alpha \sum_i \left(x_{ij} - x_{ik}\right)^2\right\} \tag{14}$$

to which the SVD can be applied as is, and we can obtain an expression similar to Eq. (13) with distinct $u_{\ell j}, \lambda_\ell, v_{\ell k}$s.

Application to various data sets

The KTD-based unsupervised FE was applied to various datasets (Fig. 1).

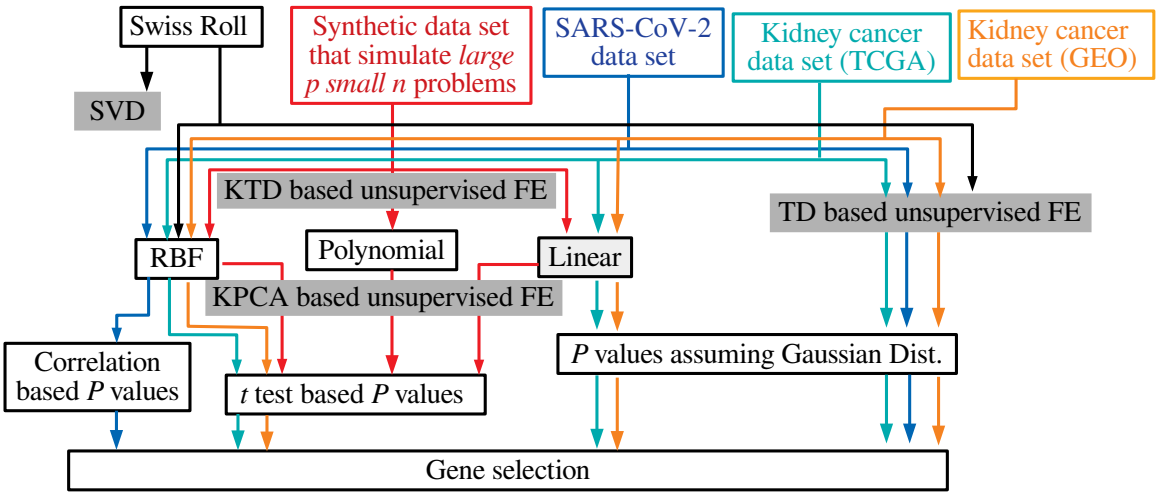


Figure 1: Overview of performed analyses

Swiss Roll

To verify if the TD extended to the kernel (KTD approach) can capture the nonlinearity, the Swiss Roll framework is considered, represented as $x_{ijk} \in \mathbb{R}^{N\times 3\times 10}$

$$p_i = -1 + \frac{2i}{N}, 1 \le i \le N \tag{15}$$

$$x_{i1k} = p_i \cos\left(2\pi p_i\right) \tag{16}$$

$$x_{i2k} = -1 + \varepsilon_{ijk} \tag{17}$$

$$x_{i3k} = p_i \sin\left(2\pi p_i\right) \tag{18}$$

where ε_{ijk} is drawn the from uniform distribution between -1 and 1 (x_{ij1} is shown in Fig. 2(A)). Here, $\sum_j x_{ijk} = 0, \sum_i x_{ijk} = N, N = 1000$. This expression is equivalent to that of 10 ensembles of Swiss Rolls, each of which are generated with distinct ε_{ijk}. To verify that the linear method cannot capture nonlinear structures (i.e., the order of i along the curved space) of the Swiss Roll, we apply the SVD to the x_{ij1} shown in Fig. 2(A). Figure 2(B) shows $u_{\ell i}, 1 \le \ell \le 3$. Clearly, the SVD cannot capture the nonlinear structure of the Swiss Roll. Moreover, even when the HOSVD is applied to x_{ijk}, the nonlinear structure is not well captured (Fig. 2(C)). The kernel based tensor can be generated as

$$x_{iki'k'} = \exp\left\{-\alpha \sum_j \left(x_{ijk} - x_{i'jk'}\right)^2\right\}, \ \alpha = 10^{-2} \tag{19}$$

to which the HOSVD can be applied to yield

$$x_{iki'k'} = \sum_{\ell_1}\sum_{\ell_3}\sum_{\ell'_1}\sum_{\ell'_3} G(\ell_1\ell_3\ell'_1\ell'_3) u_{\ell_1 i} u_{\ell_3 k} u_{\ell'_1 i'} u_{\ell'_3 k'}. \tag{20}$$

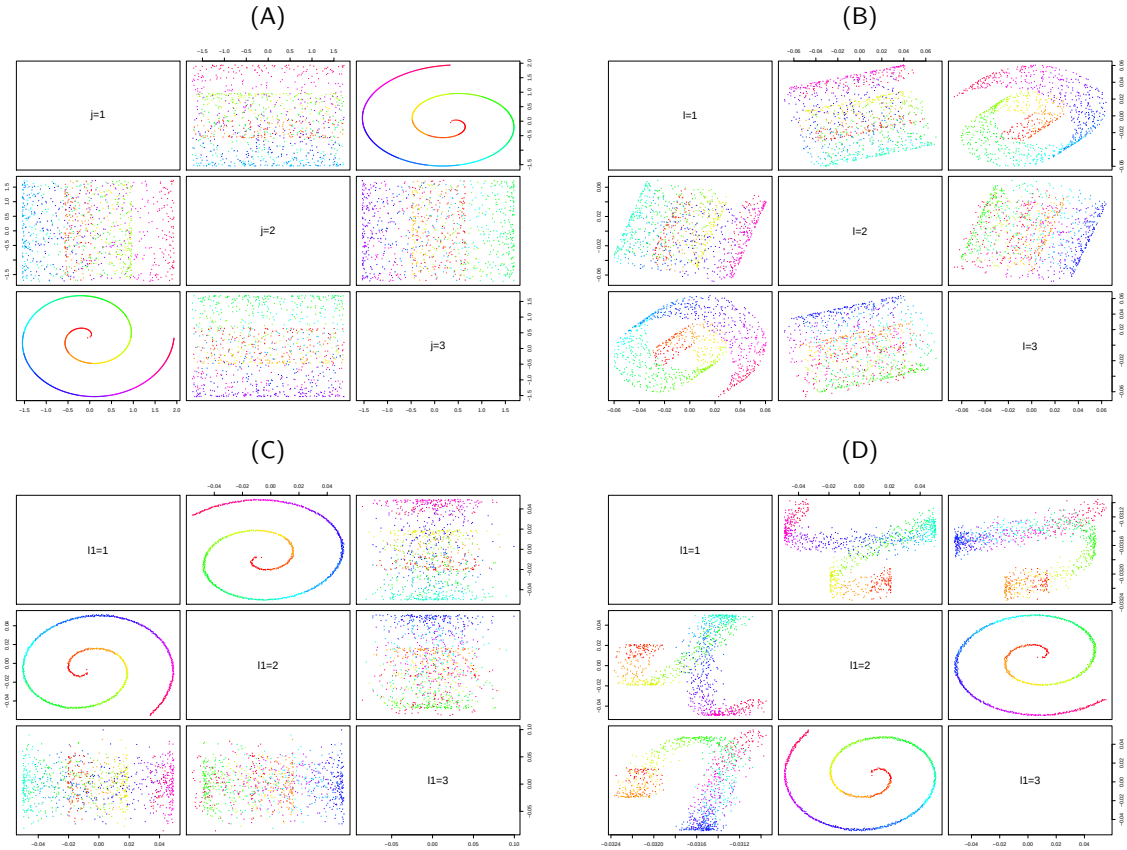


Figure 2: Swiss roll. The colors represent the direction of i (1 to N) in gradation. (A) x_{ij1}, (B) $u_{\ell i}, 1 \leq \ell \leq 3$ by SVD, (C) $u_{\ell_1 i}, 1 \leq \ell \leq 3$ by HOSVD, (D) $u_{\ell_1 i}, 1 \leq \ell \leq 3$ by kernel (RBF) based HOSVD.

$u_{\ell_1 i}, 1 \leq \ell_1 \leq 3$ represents the nonlinear structure to a certain extent (Fig. 2(D)). Thus, by replacing the inner product of the tensor with kernels, the developed TD can identify the nonlinear structure of the Swiss Roll, at least partially.

Although in the aforementioned synthetic example, the KTD-based unsupervised FE can successfully identify the nonlinear structure of Swiss Roll, which cannot be captured by the original (linear) TD-based unsupervised FE, the key intent of the TD-based unsupervised FE is to realize the feature selection in *large p small n* problems, with $p \gg n$. Thus, we must examine if the KTD-based unsupervised FE can outperform the TD-based unsupervised FE in *large p small n* problems.

Large p small n problem

We consider the following synthetic example: $x_{ijk} \in \mathbb{R}^{N \times M \times M}$, as

$$x_{ijk} \sim \begin{cases} \mathcal{N}(\mu, \sigma) & i \leq N_1 \leq N, j, k \leq \frac{M}{2} \\ \mathcal{N}(0, \sigma) & \text{otherwise} \end{cases} \tag{21}$$

where $\mathcal{N}(\mu, \sigma)$ is the Gaussian distribution of the mean μ and standard deviation σ. This problem is slightly challenging as it is a two-class problem although it appears to be a four-class problem. We apply the KTD as well as the kernel PCA (KPCA) as

$$K(x_{ijk}, x_{ij'k'}) = \sum_i x_{ijk} x_{ij'k'} \tag{22}$$

$$= \exp\left\{ -\alpha \sum_i \left(x_{ijk} - x_{ij'k'} \right)^2 \right\} \tag{23}$$

Table 1
Geometric mean P-values computed through the t tests performed on singular-value vectors determined using the KTD and KPCA. Smaller P-values are better. $N = 1000, N_1 = 10, M = 6, \mu = 2, \sigma = 1$.

Kernel	RBF ($\alpha = 10^{-6}$)		RBF ($\alpha = 10^{-3}$)		linear	
P-values	raw	corrected	raw	corrected	raw	corrected
KTD	8.19×10^{-3}	4.31×10^{-2}	1.95×10^{-2}	9.22×10^{-2}	7.18×10^{-3}	3.87×10^{-2}
KPCA	9.45×10^{-3}	2.59×10^{-1}	1.45×10^{-2}	3.89×10^{-1}	7.30×10^{-3}	2.01×10^{-1}
Kernel	polynomial ($d = 2$)		polynomial ($d = 3$)			
P-values	raw	corrected	raw	corrected		
KTD	1.36×10^{-1}	4.38×10^{-1}	3.03×10^{-1}	3.61×10^{-1}		
KPCA	7.43×10^{-2}	6.43×10^{-1}	2.28×10^{-1}	3.81×10^{-1}		

$$= \left(1 + \sum_i x_{ijk} x_{ij'k'}\right)^d \tag{24}$$

The expressions correspond to the linear, RBF, and polynomial kernels (from top to bottom). When the KPCA is applied, a tensor, x_{ijk}, is unfolded to a matrix, $x_{i(jk)} \in \mathbb{R}^{N \times M^2}$.

The performance evaluation is realized as follows. For the KTD, $u_{\ell j} u_{\ell k}, 1 \leq \ell \leq M$, are divided into two classes, $j, k \leq \frac{M}{2}$, or others, for each ℓ. A two way t test is performed on these classes. The computed M P-values for each ℓ are corrected considering the BH criterion (Taguchi, 2020), and the smallest P-value is recorded. This process is repeated 100 times, while generating new random variables from the Gaussian distribution. The geometric mean of the P-value is computed as a performance measure. For the KPCA, $u_{\ell(jk)}, 1 \leq \ell \leq M^2$ are divided into two classes for each ℓ, and the computed M^2 P-values for each ℓ are corrected using BH criterion. The smallest P-value is recorded. This process is repeated 100 times, while generating new random variables from the Gaussian distribution used to generate x_{ijk}. The geometric mean of the P-value is computed as a performance measure. Table 1 presents the results of this analysis. It can be noted that the raw P-values for the KPCA and KTD are not considerably different; nevertheless, the corrected P-values are smaller in the KTD than KPCA for the RBF and linear kernel. In other words, the KTD can effectively identify the singular-value vectors coincident in the two classes with a smaller number of singular-value vectors. Nevertheless, the RBF kernel could not outperform the linear kernel. We implemented other α values for the RBF kernel, but the RBF kernel could not outperform the linear kernel through any of the α values. This finding suggests that in *large p small n* situations, the introduction of nonlinearity in the kernel is not entirely beneficial, which is likely why the TD-based unsupervised FE could outperform the conventional statistical methods despite its linearity. Furthermore, the introduction of the nonlinearity in the kernels could not improve the performance in *large p small n* situations even when the KTD-based unsupervised FE was applied to real problems, as described in the following sections.

We also examined if the restriction of $u_{\ell(jk)}$ to $1 \leq \ell \leq M$ could improve the performance of the KPCA owing to the smaller number of P-values considered in the results. The performance was not improved, indicating that the smallest P-values for the KPCA lie in $\ell > M$; thus, $u_{\ell(jk)}, M < \ell \leq M^2$ cannot be neglected.

SARS-CoV-2 data set

We applied the KTD-based unsupervised FE to real data sets. The first dataset corresponded to the repurposing of drugs for COVID-19, which the TD-based unsupervised FE was successfully applied to the gene expression profiles of SARS-CoV-2 infected cell lines (Blanco-Melo et al., 2020; Taguchi and Turki, 2020). The TD-based unsupervised FE could predict many promising drugs including ivermectin, the clinical trials of which have been recently initiated. Herein, we briefly summarize the process implemented in the previous work (Taguchi and Turki, 2020) to enable a comparative analysis of the results of the KTD-based unsupervised FE with the previous results. The gene expression profiles were formatted as a tensor, $x_{ijkm} \in \mathbb{R}^{N \times 5 \times 2 \times 3}$, which indicated whether the gene expression of the ith gene of the jth cell line was infected ($k = 1$) or not ($k = 2$, control), considering three biological replicates. The HOSVD was applied to x_{ijkm} to yield

$$x_{ijkm} = \sum_{\ell_1=1}^{5} \sum_{\ell_2=1}^{2} \sum_{\ell_3=1}^{m} \sum_{\ell_4=1}^{N} G(\ell_1 \ell_2 \ell_3 \ell_4) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell_4 i} \tag{25}$$

Table 2
Confusion matrix of the selected genes for the KTD and previous study (Taguchi and Turki, 2020). In the KTD-based unsupervised FE, the ranking was based on the correlation coefficients.

		KTD-based unsupervised FE rank					
		$\alpha = 10^{-2}$		$\alpha = 10^{-4}$		$\alpha = 10^{-6}$	
		> 163	≤ 163	> 163	≤ 163	> 163	≤ 163
Previous study	adjusted P-values ≥ 0.01	21516	118	21509	125	21540	94
	adjusted P-values < 0.01	118	45	125	38	94	69
Fisherś exact test	P-value	5.05×10^{-59}		1.53×10^{-46}		6.61×10^{-108}	
	odds ratio	69		52		168	

where $G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{5\times2\times3\times N}$ is a core tensor, and $u_{\ell_1 j} \in \mathbb{R}^{5\times5}, u_{\ell_2 k} \in \mathbb{R}^{2\times2}, u_{\ell_3 m} \in \mathbb{R}^{3\times3}, u_{\ell_4 i} \in \mathbb{R}^{N\times N}$ are orthogonal singular value matrices.

The purpose of the analysis was to identify the genes whose expression was distinct between the control and infected cells, independent of the cell lines and replicates. To this end, we selected ℓ_1, ℓ_2, ℓ_3 with constant $u_{\ell_1 j}$s and $u_{\ell_3 m}$s and $u_{\ell_2 1} = -u_{\ell_2 2}$. It was noted that $\ell_1 = \ell_3 = 1$ and $\ell_2 = 2$ could satisfy these requirements. Next, we identified $\ell_4 = 5$ associated with G having the largest absolute values given $\ell_1 = 1, \ell_2 = 2, \ell_3 = 1$. Once $\ell_4 = 5$ was selected, the P-values were assigned to gene i, assuming the null hypothesis that $u_{\ell_4 i}$ obeys the Gaussian distribution:

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{5i}}{\sigma_5}\right)^2\right] \tag{26}$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution in which the argument is larger than x, and σ_5 is the standard deviation. The obtained P-values were corrected using the BH criterion (Taguchi, 2020). A total of 163 genes for which the adjusted P-values were less than 0.01 were selected and used to predict the drugs effective against COVID-19.

The objective of this study was to compare the performance of the KTD-based unsupervised FE applied to x_{ijkm} with that of an existing study (Taguchi and Turki, 2020). Therefore, we employed the RBF kernel as

$$x_{jkmj'k'm'} = K(x_{ijkm}, x_{ij'k'm'}) = \exp\left\{-\alpha \sum_i \left(x_{ijkm} - x_{ij'k'm'}\right)^2\right\} \tag{27}$$

with $\alpha = 1 \times 10^{-2}, 1 \times 10^{-4}, 1 \times 10^{-6}$, to which the HOSVD was applied. We obtained

$$x_{jkmj'k'm'} = \sum_{\ell_1=1}^{5}\sum_{\ell_2=1}^{2}\sum_{\ell_3=1}^{3}\sum_{\ell'_1=1}^{5}\sum_{\ell'_2=1}^{2}\sum_{\ell'_3=1}^{3} G(\ell_1\ell_2\ell_3\ell'_1\ell'_2\ell'_3)u_{\ell_1 j}u_{\ell_2 k}u_{\ell_3 m}u_{\ell'_1 j'}u_{\ell'_2 k'}u_{\ell'_3 m'} \tag{28}$$

It was observed that u_{1j} and u_{1m} were constant, and $u_{\ell_2 1} = -u_{\ell_2 2}$ with $\ell_2 = 2$. We compared the $u_{1j}u_{2k}u_{1m}$ computed using TD-based unsupervised FE (Taguchi and Turki, 2020) and KTD with k. It was clarified that $u_{1j}u_{2k}u_{1m}$ computed using the KTD were more notably coincident with k (Fig. 3). Thus, the KTD exhibited a slight improvement over the TD-based unsupervised FE (Taguchi and Turki, 2020). Next, we were required to select genes, although the process was challenging as $u_{\ell_4 i}$ used for gene selection could not be obtained. To apply the KTD for gene selection, we recomputed $u_{1j}u_{2k}u_{1m}$ excluding i sequentially. In addition, the correlation coefficients between $u_{1j}u_{2k}u_{1m}$ and k were obtained. Subsequently, the is were ranked in ascending order of the absolute values of the correlation coefficients. Excluding the genes with expressions coincident with the distinction between $k = 1$ and $k = 2$ was expected to result in less significant (i.e., smaller absolute values of) correlation coefficients. Table 2 presents the confusion matrix of the selected genes between the previous study (Taguchi and Turki, 2020) and KTD. Notably, the genes are highly coincident and can even be regarded as identical considering that there exists 20,000 genes, of which we selected only 163.

Next, we biologically evaluated the 163 genes selected using the KTD-based unsupervised FE. In particular, we compared the selected 163 genes with gold standard human proteins that are known to interact with the SARS-CoV-2

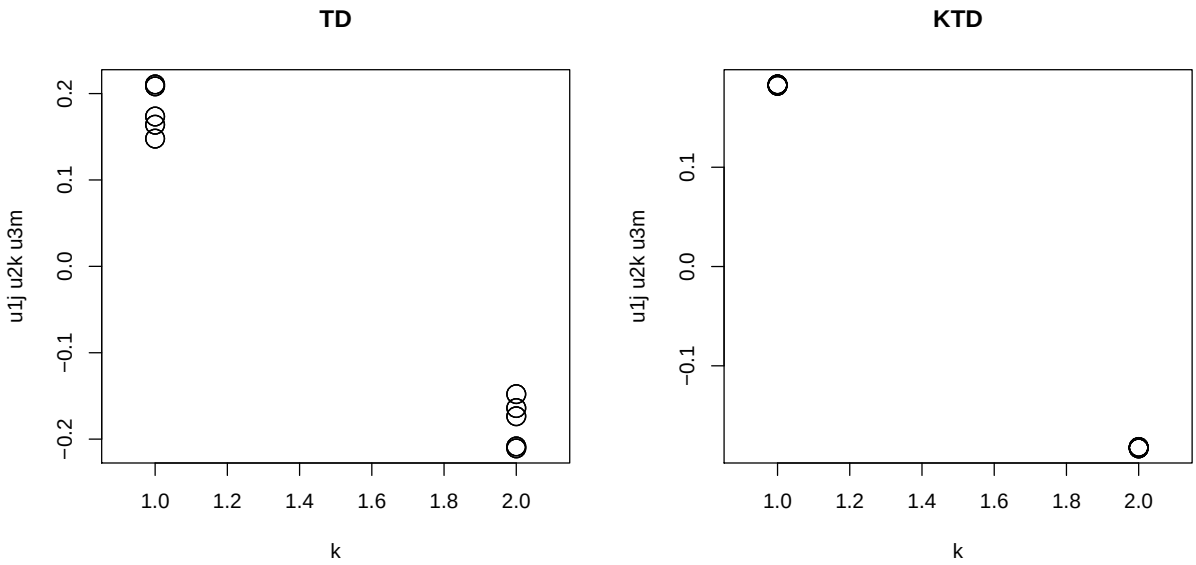


Figure 3: Scatter plot of $u_{1j}u_{2k}u_{1m}$ and k. left:TD (Taguchi and Turki, 2020), right:KTD ($\alpha = 1 \times 10^{-6}$).

proteins during infection (Gordon et al., 2020), following the procedure described in the previous study (Taguchi and Turki, 2020) (Table 3).

The 163 genes selected by the KTD-based unsupervised FE are highly coincident with the gold standard SARS-CoV-2 interacting human proteins, the results in Table 3 even outperformed the results for the gold standard SARS-CoV-2 interacting human proteins compared to the results presented in Table S33 (Taguchi and Turki, 2020) corresponding to the application of the TD-based unsupervised FE (Taguchi and Turki, 2020). Notably, all 27 protein groups when $\alpha = 10^{-2}$ exhibit larger odds ratio as well as smaller P-values than those in the previous study (corresponding to the TD-based unsupervised FE (Taguchi and Turki, 2020)) in Table 3.

Tables 3 indicates that the KTD based unsupervised FE outperformed the TD-based unsupervised FE when it was applied to a real data set in contrast to when the KTD-based unsupervised FE was applied to synthetic data simulating a *large p small n* problem (eq.(21)). This is possibly because synthetic data set could not fully emulate complicated structures in real data sets; KTD-based unsupervised FE should be tested with more real data sets to test whether it can outperform TD-based unsupervised FE.

Kidney cancer data sets

In a previous work, we attempted to realize the integrated analysis of multiomics data by applying the TD-based unsupervised FE (Ng and Taguchi, 2020). In the existing study (Ng and Taguchi, 2020), the objective was to select genes that act as prognostic biomarker for kidney cancer. To this end, the TD-based unsupervised FE was employed to identify the differentially expressed genes (DEGs) between normal kidneys and tumors in two independent data sets. The TD-based unsupervised FE successfully identified 72 mRNAs and 11 miRNAs for the first data set retrieved from the TCGA, and 209 mRNAs and 3 miRNAs for the second data set retrieved from the GEO. To extend this analysis, in the present study, the mRNA expression matrix $x_{ij} \in \mathbb{R}^{N \times M}$, which represents the ith mRNA expression in the jth sample and the miRNA expression matrix $x_{kj} \in \mathbb{R}^{K \times M}$, which represents the kth miRNA expression in the jth sample were integrated into the tensor $x_{ijk} \in \mathbb{R}^{N \times M \times K}$ as

$$x_{ijk} = x_{ij} x_{kj} \tag{29}$$

and converted to a matrix $x_{ik} \in \mathbb{R}^{K \times M}$

$$x_{ik} = \sum_j x_{ijk} \tag{30}$$

Table 3
Coincidence between 163 genes and human proteins that are known to interact with SARS-CoV-2 proteins during infection. The P-values were computed by applying Fisher's exact tests to the confusion matrix. For more details, please refer to the reference (Taguchi and Turki, 2020).

SARS-CoV-2 proteins	P values				Odds Ratio			
	previous study	RBF kernel $\alpha = 10^{-2}$	10^{-4}	10^{-6}	previous study	RBF kernel $\alpha = 10^{-2}$	10^{-4}	10^{-6}
SARS-CoV2 E	6.55×10^{-27}	1.08×10^{-44}	4.90×10^{-22}	6.58×10^{-26}	10.16	16.18	8.63	9.84
SARS-CoV2 M	1.37×10^{-26}	1.66×10^{-37}	4.14×10^{-22}	4.14×10^{-22}	8.41	11.40	7.26	7.26
SARS-CoV2 N	4.61×10^{-24}	7.22×10^{-38}	3.51×10^{-15}	1.46×10^{-29}	11.42	17.11	7.93	13.63
SARS-CoV2 nsp1	1.05×10^{-20}	3.98×10^{-28}	4.26×10^{-14}	5.51×10^{-15}	10.00	12.92	7.43	7.78
SARS-CoV2 nsp10	3.40×10^{-20}	9.04×10^{-26}	4.44×10^{-17}	3.88×10^{-19}	11.51	14.13	10.06	11.02
SARS-CoV2 nsp11	1.13×10^{-29}	2.62×10^{-38}	4.00×10^{-20}	1.2×10^{-27}	10.66	13.43	7.74	10.03
SARS-CoV2 nsp12	4.87×10^{-20}	1.64×10^{-29}	1.85×10^{-14}	2.41×10^{-15}	9.48	13.09	7.38	7.72
SARS-CoV2 nsp13	6.04×10^{-33}	3.08×10^{-37}	2.19×10^{-20}	7.41×10^{-28}	11.17	12.49	7.51	9.65
SARS-CoV2 nsp14	1.75×10^{-22}	1.08×10^{-31}	2.72×10^{-16}	2.78×10^{-18}	12.05	16.26	9.30	10.18
SARS-CoV2 nsp15	1.85×10^{-20}	4.26×10^{-29}	1.71×10^{-17}	1.08×10^{-14}	10.23	13.75	9.03	7.90
SARS-CoV2 nsp2	4.81×10^{-33}	5.16×10^{-42}	3.31×10^{-20}	6.34×10^{-31}	11.79	14.76	7.83	11.11
SARS-CoV2 nsp4	5.79×10^{-29}	5.25×10^{-42}	4.09×10^{-16}	5.53×10^{-26}	10.26	14.41	6.47	9.36
SARS-CoV2 nsp5	3.78×10^{-25}	4.18×10^{-38}	6.97×10^{-18}	4.63×10^{-24}	12.36	17.97	9.36	11.91
SARS-CoV2 nsp5^{+}	3.75×10^{-17}	1.42×10^{-21}	3.74×10^{-17}	4.65×10^{-14}	11.39	13.78	11.39	9.72
SARS-CoV2 nsp6	9.47×10^{-26}	7.15×10^{-35}	3.91×10^{-21}	4.97×10^{-22}	9.00	11.71	7.68	7.94
SARS-CoV2 nsp7	1.93×10^{-29}	1.09×10^{-33}	2.70×10^{-16}	1.61×10^{-22}	10.81	12.18	6.76	8.65
SARS-CoV2 nsp8	1.11×10^{-29}	1.68×10^{-41}	1.57×10^{-18}	1.11×10^{-28}	10.14	13.80	6.95	9.84
SARS-CoV2 nsp9	5.36×10^{-29}	1.46×10^{-40}	8.17×10^{-19}	6.71×10^{-28}	12.24	16.67	8.59	11.85
SARS-CoV2 orf10	5.29×10^{-34}	2.44×10^{-44}	5.57×10^{-20}	1.16×10^{-26}	12.37	15.95	7.94	10.01
SARS-CoV2 orf3a	2.06×10^{-28}	1.31×10^{-43}	3.17×10^{-18}	1.62×10^{-24}	9.95	14.76	6.99	8.80
SARS-CoV2 orf3b	1.89×10^{-29}	7.58×10^{-41}	1.35×10^{-17}	2.50×10^{-27}	11.80	15.94	7.78	11.06
SARS-CoV2 orf6	8.81×10^{-26}	5.67×10^{-39}	5.22×10^{-13}	8.03×10^{-22}	10.37	14.99	6.14	9.05
SARS-CoV2 orf7a	1.69×10^{-28}	9.79×10^{-35}	8.26×10^{-22}	1.18×10^{-23}	10.00	11.90	8.04	8.57
SARS-CoV2 orf8	5.94×10^{-28}	2.48×10^{-39}	4.32×10^{-18}	1.27×10^{-20}	9.25	12.57	6.57	7.25
SARS-CoV2 orf9b	6.54×10^{-30}	7.07×10^{-44}	5.12×10^{-17}	9.11×10^{-28}	12.12	17.34	7.68	11.36
SARS-CoV2 orf9c	1.11×10^{-28}	1.03×10^{-42}	1.15×10^{-21}	5.76×10^{-26}	8.35	12.07	6.66	7.69
SARS-CoV2 Spike	8.22×10^{-26}	3.48×10^{-41}	8.46×10^{-15}	6.07×10^{-22}	10.08	15.34	6.55	8.81

+ with mutation, C145A

to which the SVD was applied to yield

$$x_{ik} = \sum_{\ell} u_{\ell i} \lambda_{\ell} v_{\ell k}. \tag{31}$$

The missing singular-value vectors attributed to the jth mRNA and miRNA samples were recovered as follows:

$$u^{\text{mRNA}}_{\ell j} = \sum_{i} x_{ij} u_{\ell i} \tag{32}$$

$$u^{\text{miRNA}}_{\ell j} = \sum_{k} x_{kj} u_{\ell k} \tag{33}$$

The P-values attributed to mRNAs and miRNAs were derived as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{2i}}{\sigma_2} \right)^2 \right] \tag{34}$$

$$P_k = P_{\chi^2}\left[> \left(\frac{v_{2k}}{\sigma'_2} \right)^2 \right] \tag{35}$$

as u_{2j}^{mRNA} and u_{2j}^{miRNA} were noted to be coincident with the distinction between the tumors and normal kidneys. The mRNAs and miRNAs with the adjusted P-values of less than 0.01 were selected.

Notably, the kernel trick cannot be applied to this formulation as the summation is obtained over the sample index j and not the feature index i, k. Thus, instead, we formulated a kernel-trick-friendly integration of x_{ij} and x_{kj} as

$$x_{jj'}^{\text{mRNA}} = K(x_{ij}, x_{ij'}) = \sum_i x_{ij} x_{ij'} \tag{36}$$

$$x_{jj'}^{\text{miRNA}} = K(x_{kj}, x_{kj'}) = \sum_k x_{kj} x_{kj'} \tag{37}$$

$$x_{jj'} = \sum_{j''} x_{jj''}^{\text{mRNA}} x_{j''j'}^{\text{miRNA}} \tag{38}$$

to which the SVD can be applied to yield

$$x_{jj'} = \sum_\ell u_{\ell j} \lambda v_{\ell_{j'}}. \tag{39}$$

As this expression includes the inner product, for which the summation is obtained over the feature index i (eq. (36)) and k (eq. (37)), the expressions can be easily extended to the RBF kernel as

$$x_{jj'}^{\text{mRNA}} = K(x_{ij}, x_{ij'}) = \exp\left\{-\alpha \sum_i \left(x_{ij} - x_{ij'}\right)^2\right\} \tag{40}$$

$$x_{jj'}^{\text{miRNA}} = K(x_{kj}, x_{kj'}) = \exp\left\{-\alpha' \sum_k \left(x_{kj} - x_{kj'}\right)^2\right\}. \tag{41}$$

$x_{jj'}$ was computed using Eq. (38), and the SVD was applied to $x_{jj'}$ to obtain an expression similar to Eq. (39).

A possible biological validation of the performances of the KTD-based unsupervised FE is by evaluating $u_{\ell j}$ and $v_{\ell j}$, corresponding to the mRNA and miRNA samples, respectively. Specifically, we consider whether

- $u_{\ell j}$s are distinct between tumorous and normal kidneys,
- $v_{\ell j}$s are distinct between tumorous and normal kidneys,
- $u_{\ell j}$s and $v_{\ell j}$s are coincident.

To examine the first point, we compute $u_{\ell j}$ and $v_{\ell j}$ using linear (eqs. (36) and (37)) and RBF (eqs. (40) and (41)) kernels. As in the previous study (Ng and Taguchi, 2020), the second singular-value vectors, u_{2j} and v_{2j}, are noted to be the most coincident with the aforementioned three requirements. Table 4 presents the comparisons of the P-values used to evaluate the aforementioned three conditions for the linear kernel, RBF kernel, and the approach employed in the previous study. The P-values to evaluate the distinction between tumorous and normal kidneys were computed through the t test, whereas those evaluating the coincidence between u_{2j} and v_{2j} were computed by considering the correlation coefficients. Although the three types of approaches exhibited a reasonable performance, the RBF kernel that considered the nonlinearity could not always outperform the two other linear methods. The superiority of the RBF kernel was only noted occasionally depending on α and α'. Thus, the nonlinear method could not always outperform the linear methods in the *large p small n* situation.

Finally, we compared the genes selected using the three methods. For the RBF, we followed the procedure pertaining to the RBF kernel TD-based unsupervised FE applied to the SARS-CoV-2 infection cell lines. $x_{jj'}$ was computed excluding one of the miRNAs or mRNAs, and the SVD was applied to the obtained $x_{jj'}$ to obtain u_{2j}. The t test was applied to u_{2j} to compute the P-values using which the distinction of u_{2j} between tumorous and normal kidneys could be evaluated. When the first data set was considered, 72 top ranked mRNAs and 11 top ranked miRNAs having large (thus, less significant) P-values were selected because the exclusion of the mRNAs or miRNAs distinct between the tumorous and normal kidneys was expected to decrease the significance of the distinct u_{2j} between the two entities. Similarly, when the second data set was considered, 209 top ranked mRNAs and 3 top ranked miRNAs were selected.

Table 4
Performances of the linear and RBF kernel TD-based unsupervised FE and that achieved in the previous study (Ng and Taguchi, 2020) by using Eq. (30). The P-values marked by an asterisk correspond to the most significant values.

	tumor vs. normal kidney		correlation coefficient
	u_{2j} (mRNA)	v_{2j} (miRNA)	u_{2j} vs. v_{2j}
1st data set (TCGA)			
linear kernel	1.27×10^{-38}	1.76×10^{-62}	0.912 (1.28×10^{-126})
RBF kernel ($\alpha = 10^{-10}$ and $\alpha' = 10^{0}$)	1.99×10^{-14}	1.99×10^{-14}	1.00 (0.00*)
RBF kernel ($\alpha = 10^{-12}$ and $\alpha' = 10^{-2}$)	1.06×10^{-44}*	4.24×10^{-38}	0.974 (5.93×10^{-212})
RBF kernel ($\alpha = 10^{-14}$ and $\alpha' = 10^{-4}$)	4.70×10^{-43}	2.89×10^{-66}	0.914 (2.00×10^{-128})
previous study	7.10×10^{-39}	2.13×10^{-71}*	0.905 (1.63×10^{-121})
2nd data set (GEO)			
linear kernel	5.36×10^{-23}*	2.76×10^{-18}	0.937 (3.91×10^{-16})
RBF kernel ($\alpha = 10^{-4}$ and $\alpha' = 10^{0}$)	5.61×10^{-7}	5.61×10^{-7}	1.00 (0.00*)
RBF kernel ($\alpha = 10^{-6}$ and $\alpha' = 10^{-2}$)	5.42×10^{-15}	1.17×10^{-13}	0.966 (2.06×10^{-20})
RBF kernel ($\alpha = 10^{-8}$ and $\alpha' = 10^{-4}$)	6.93×10^{-18}	2.93×10^{-15}	0.921 (1.16×10^{-14})
previous study	6.74×10^{-22}	2.54×10^{-18}*	0.931 (1.58×10^{-15})

Table 5
Confusion matrix of the selected genes between the KTD (linear and RBF kernel) and TD-based unsupervised FE (previous study (Ng and Taguchi, 2020) using Eq. (30)). The ranking in the KTD-based unsupervised FE was t test based.

1st data set (TCGA)		linear kernel adjusted P		RBF kernel rank					
				$(\alpha, \alpha') = (10^{-10}, 10^{0})$		$(\alpha, \alpha') = (10^{-12}, 10^{-2})$		$(\alpha, \alpha') = (10^{-14}, 10^{-4})$	
mRNA		> 0.01	< 0.01	> 72	≤ 72	> 72	≤ 72	> 72	≤ 72
previous study	> 0.01	19458	6	19407	57	19437	27	19439	25
adjusted P	< 0.01	7	65	57	15	27	45	25	47
miRNA		> 0.01	< 0.01	> 11	≤ 11	> 11	≤ 11	> 11	≤ 11
previous study	> 0.01	814	0	—	—	804	10	807	7
adjusted P	< 0.01	0	11	—	—	10	1	7	4
2nd data set (GEO)		linear kernel adjusted P		RBF kernel rank					
				$(\alpha, \alpha') = (10^{-4}, 10^{0})$		$(\alpha, \alpha') = (10^{-6}, 10^{-2})$		$(\alpha, \alpha') = (10^{-8}, 10^{-4})$	
mRNA		> 0.01	< 0.01	> 209	≤ 209	> 209	≤ 209	> 209	≤ 209
previous study	> 0.01	33787	2	33583	206	33620	169	33604	185
adjusted P	< 0.01	56	153	206	3	169	40	185	24
miRNA		> 0.05	< 0.05	> 3	≤ 3	> 3	≤ 3	> 3	≤ 3
previous study	> 0.05	316	0	—	—	314	2	315	1
adjusted P	< 0.05	0	3	—	—	2	1	1	2

For the linear kernel, the P-values attributed to the miRNAs and miRNAs were computed using Eqs. (34) and (35) with

$$u_{2i} = \sum_j x_{ij} u_{2j} \tag{42}$$

$$v_{2k} = \sum_j x_{kj} v_{2j}. \tag{43}$$

Subsequently, the mRNAs and miRNAs with adjusted P-values of less than 0.01 were selected. Table 5 presents a comparison of the confusion matrices of the selected genes between the present study and previous study (Ng and Taguchi, 2020). The higher coincidence of the RBF or linear kernel with that obtained in the previous study depended on α and α'. The RBF kernels were more coincident with the findings of the previous study for smaller α and α', although liner kernel was always more coincident with the previous study where TD-based unsupervised FE was employed.

Although no gold standard data sets exist to evaluate the selected genes, one possible evaluation pertains to the

coincidence of the selected genes between the first (TCGA) and second data sets (GEO). As these two data sets are independent, it is unlikely that commonly selected genes are present, and the selected genes are at most 1% of all genes. In the previous study (Ng and Taguchi, 2020) that employed Eq. (30), 11 genes were commonly selected between TCGA and GEO, whereas as small as 7 genes were commonly selected between the TCGA and GEO for the linear and RBF kernels ($\alpha = 10^{-12}$ and $\alpha' = 10^{-2}$ that achieved the maximum overlap among the tested RBF kernels). Although these overlaps are still significant (using Fisher's exact test, $P = 3.43 \times 10^{-6}$, odds ratio: 12.54 for RBF kernel and $P = 7.93 \times 10^{-7}$, odds ratio: 15.82 for linear kernel), the values are inferior to those for the previous study (Ng and Taguchi, 2020) that involved 11 commonly selected genes ($P = 8.97 \times 10^{-11}$, odds ratio: 19.7). Thus, the KTD-based unsupervised FEs are at most comparable with the TD-based unsupervised FE in this *large p small n* problem.

Discussion

In the aforementioned analyses, we successfully formalized the KTD-based unsupervised FE to be applied to select DEGs. Although the implementation was effective, the KTD-based unsupervised FE could not always outperform the TD based unsupervised FE that address only the linearities. This finding may appear to be unexpected, as nonlinear methods generally outperform linear methods pertaining to nonlinear application by definition. Nevertheless, this expectation does not always hold, especially when the linear methods can achieve optimal solutions similar to those attained by nonlinear methods.

As shown in Fig. 2(D), the KTD can precisely recognize the nonlinearity in Swiss Rolls, at least partially. However, when the KTD-based unsupervised FE is applied to *large p small n* problems, even the use of nonlinear kernels could not always outperform the TD-based unsupervised FE or linear kernel based KTD-based unsupervised FE. The reason for this aspect can be explained as follows. To represent complicated structures that only nonlinear methods can recognize, many points are required. For example, to represent nth order polynomials, we need at least $n + 1$ points. As the higher order polynomials can represent more complicated structures, more points are necessary to realize this representation. In *large p small n* problems, the number of samples is insufficient to represent complicated structures. As shown in Fig. 2, there exist 1,000 points in 3-dimensional space. In contrast, in the considered *large p small n* problems, the number of samples ranges from a few tens to hundreds, whereas the number of corresponding features varies from 10^2 to 10^4. Thus, clearly, the points cannot effectively represent complicated structures. The TD-based unsupervised FE, despite its linearity, has outperformed conventional statistical test-based gene selection methods. It is rather unexpected that such a simple linear method can outperform more advanced tools. Specifically, the TD-based unsupervised FE has already achieved the performances that KTD-based unsupervised FE employing nonlinear kernels is expected to achieve in *large p small n* problems.

In addition, the KTD-based unsupervised FE cannot exploit certain advantages that the TD-based unsupervised FE exhibits. For example, as the KTD cannot yield singular-value vectors attributed to the features (genes), we cannot apply the empirical P-value computation assuming that the singular-value vectors obey the Gaussian distribution. Instead, we must repeatedly compute the singular-value vectors excluding the features sequentially, which is a time-consuming process. In addition, the computation of inner products exponentially doubles the amount of memory required. As the number of samples is small, this aspect may not be an immediate problem. Nevertheless, this process for the first data set (TCGA) required half a day of CPU time, whereas using 12 CPU units as the set included hundreds of samples. In contrast, the TD-based unsupervised FE can complete the corresponding computation in a few minutes. Furthermore, the P-values computed by the KTD-based unsupervised FE cannot be directly used for gene selection. When used to rank the genes, the process is effective (Tables 2 and 5). However, the process is ineffective when used to screen features directly because the P-values are sufficiently small even after excluding one variable feature, e.g., those distinct between two classes (e.g., control and cancers). In this scenario, we cannot identify the number of features to be selected by only considering the P-values. Nevertheless, this decision can be made using TD-based unsupervised FE.

Despite these disadvantages of KTD-based unsupervised FE, there may exist scenarios in which the results can be improved, whereas this aspect does not hold true for TD-based unsupervised FE that lacks tunable parameters. At present, although we were unable to identify the cases in which the KTD unsupervised FE employing nonlinear kernels could outperform the TD-based unsupervised FE in *large p small n* problems, it is expected that the proposed approach can outperform the existing approach in scenarios involving *large p small n* problems in which the TD-based unsupervised FE cannot achieve a high performance.

The reason we reviewed neither recent approaches dealing with *large p small n* problems nor tensor learning approaches is primarily because our approach is very distinct from other approaches. First, the recent approaches that deal with *large p small n* problems are not applicable to the extreme cases that we dealt with in this study. In this study, the numbers of features (i.e., genes) are 10^4, whereas the numbers of samples are from 10 to 10^2. Thus, the ratio of the former to the latter is as large as 10^2, whereas conventional approaches regarded to be applicable to *large p small n* problems usually do not assume such extreme cases. For example, although Loh (2011) reviewed various tools that deal with *large p small n* problems, he applied the methods to at most $p = 500$ whereas $n = 100$; the number of features is only 5 times larger than that of samples, which is much less than 10^2. Chakraborty et al. (2012) applied Bayesian nonlinear regression for *large p small n* problems that have only 700 features for which as many as 62 samples were measured; thus, the ratio of the number of features to the number of samples is about 10, which is also much less than 10^2. Secchi et al. (2013) dealt with extreme case where $n = 20$ and $p = 1024$, whereas they did not apply to feature selection because their simulated data was not composed of the mixture of features with and without distinction between two classes but was composed of features distinct between two classes or those without distinction. Although there are only a few examples and it is impossible to prove that there are no methods applicable to the present data sets, we need the methods satisfying the following conditions:

- applicable to *large p small n* problems where $\frac{p}{n} > 10^2$
- need to extract generated futures that can discriminate between multiple classes
- selecting a small number of features biologically significant (e.g. enrichment analysis, see Table 3) with stability (e.g., coincidence between two independent data sets, see Table 5)

which are not likely fulfilled with the methods that are not specifically designed to be fitted to genomic science.

This difference between our approach and general machine learning solutions that aim to solve the *large p small n* problem prevents us from applying other methods to our problem. For example, fast adaptive K-means (Wang et al., 2019a) (FAKM) aims to select features coincident with clustering using K-means algorithm. When FAKM was applied to synthetic data (eq. (21)), in spite of FAKM almost always correctly selecting features $i \leq N_1$, K-means practically fails to generate two clusters coincident with the distinction between $j, k \leq \frac{M}{2}$ and others. This is because simply selecting $i \leq N_1$ features cannot allow K-means to generate correct clusters because of the too-noisy data set. Because FAKM simply tries to select features relatively coincident with clustering, it cannot guarantee the generation of the correct clustering. On the other hand, although Wang et al. (2019b) proposed methods that can generate low rank latent variables that can discriminate clusters correctly, it does not have ability to select features. To our knowledge, no other general machine learning methods applicable to the present problem exist.

As for tensor learning, the term "tensor learning" often simply means deriving low rank representations of given tensors, typically using tensor decomposition (Guo et al., 2012; Tao et al., 2007; Liu et al., 2016). Thus, the main purpose of the so-called tensor learning is to approximate the whole tensor with low rank representation. In contrast to these purposes of tensor learning, we need to find latent variables coincident with the properties of interest (e.g., distinction between infectious and control cell lines, Fig. 3), regardless of how well it can approximate a whole tensor. Thus, our purpose differs from the general purpose of tensor learning in which how well it can approximate the whole tensor is the question. From this point view, tensor learning is not deeply related to our above purpose.

As for the comparisons with the methods designed to be fitted to genomic science, superiority of previously proposed TD-based unsupervised FE over the conventional methods was repeatedly demonstrated. For example, when we previously investigated SARS-CoV-2 data set (Taguchi and Turki, 2020) analyzed in this study, three conventional methods frequently used to identify differentially expressed genes, *t* test, SAM (Tusher et al., 2001) and limma (Ritchie et al., 2015) were shown to be inferior to TD-based unsupervised FE. When we previous analyzed kidney cancer data sets (Ng and Taguchi, 2020) analyzed in this study, these three methods were shown to be inferior to TD-based unsupervised FE. Other than these two studies, superiority of TD-based unsupervised FE over conventional methods was repeatedly demonstrated. For example, when TD-based unsupervised FE was applied to gene expression profiles measured in multiple mice tissues under the multiple experimental conditions (Taguchi, 2017), categorical regression (also known as ANOVA), SAM and limma were shown to be inferior to TD-based unsupervised FE. Based on these previous studies, the superiority of TD-based unsupervised FE over conventional methods was established. Therefore, we concentrated on the comparisons between newly proposed kernel TD-based unsupervised FE and original TD-based unsupervised FE in this study; extensive comparisons with other methods than TD-based unsupervised FE was not critical.

Finally, we added more comments about cpu time used for KTD-based unsupervised FE. When liner kernel, eq. (22) was employed, cpu time was as short as TD-based unsupervised FE because we made use of equations like eqs. (42) and (43) for selecting genes. Nevertheless, when we employ RBF kernel, we need to recompute TD every time we exclude one feature. When we performed one run of RBF kernel based KTD based unsupervised FE to SARS-CoV-2 infection data set, it took 30 minutes on cpu E5-2643 3.40GHz with 256GB memory. One run of the same computation to TCGA kidney data took 300 h on the same machine. CPU time is mainly dependent upon the square of sample size. For SARS-CoV-2, there are only 30 samples, whereas for TCGA kidney data, there are 324 samples. The ratio of samples size is $324/30 \sim 10$, whose square is 10^2. The ratio of cpu time required is $300/0.5 \sim 6\times10^2$. These 2 match with each other excluding numerical factor less than 10. Thus, it is unrealistic to apply KTD-based unsupervised FE to the samples larger than 10^3.

Methods

Data set

The data sets used in this study are all public domain. The information regarding the retrieval of these data sets has been presented in previous studies (Taguchi and Turki, 2020; Ng and Taguchi, 2020), in which the TD-based unsupervised FE approach was applied to these data sets.

References

Blanco-Melo, D., Nilsson-Payant, B.E., Liu, W.C., Uhl, S., Hoagland, D., Møller, R., Jordan, T.X., Oishi, K., Panis, M., Sachs, D., Wang, T.T., Schwartz, R.E., Lim, J.K., Albrecht, R.A., tenOever, B.R., 2020. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell 181, 1036–1045.e9. URL: https://doi.org/10.1016/j.cell.2020.04.026, doi:10.1016/j.cell.2020.04.026.

Chakraborty, S., Ghosh, M., Mallick, B.K., 2012. Bayesian nonlinear regression for large p small n problems. Journal of Multivariate Analysis 108, 28 – 40. URL: http://www.sciencedirect.com/science/article/pii/S0047259X12000164, doi:https://doi.org/10.1016/j.jmva.2012.01.015.

Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., Tummino, T.A., Hüttenhain, R., Kaake, R.M., Richards, A.L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., Kim, M., Haas, P., Polacco, B.J., Braberg, H., Fabius, J.M., Eckhardt, M., Soucheray, M., Bennett, M.J., Cakir, M., McGregor, M.J., Li, Q., Meyer, B., Roesch, F., Vallet, T., Kain, A.M., Miorin, L., Moreno, E., Naing, Z.Z.C., Zhou, Y., Peng, S., Shi, Y., Zhang, Z., Shen, W., Kirby, I.T., Melnyk, J.E., Chorba, J.S., Lou, K., Dai, S.A., Barrio-Hernandez, I., Memon, D., Hernandez-Armenta, C., Lyu, J., Mathy, C.J.P., Perica, T., Pilla, K.B., Ganesan, S.J., Saltzberg, D.J., Rakesh, R., Liu, X., Rosenthal, S.B., Calviello, L., Venkataramanan, S., Liboy-Lugo, J., Lin, Y., Huang, X.P., Liu, Y., Wankowicz, S.A., Bohn, M., Safari, M., Ugur, F.S., Koh, C., Savar, N.S., Tran, Q.D., Shengjuler, D., Fletcher, S.J., O'Neal, M.C., Cai, Y., Chang, J.C.J., Broadhurst, D.J., Klippsten, S., Sharp, P.P., Wenzell, N.A., Kuzuoglu-Ozturk, D., Wang, H.Y., Trenker, R., Young, J.M., Cavero, D.A., Hiatt, J., Roth, T.L., Rathore, U., Subramanian, A., Noack, J., Hubert, M., Stroud, R.M., Frankel, A.D., Rosenberg, O.S., Verba, K.A., Agard, D.A., Ott, M., Emerman, M., Jura, N., von Zastrow, M., Verdin, E., Ashworth, A., Schwartz, O., d'Enfert, C., Mukherjee, S., Jacobson, M., Malik, H.S., Fujimori, D.G., Ideker, T., Craik, C.S., Floor, S.N., Fraser, J.S., Gross, J.D., Sali, A., Roth, B.L., Ruggero, D., Taunton, J., Kortemme, T., Beltrao, P., Vignuzzi, M., García-Sastre, A., Shokat, K.M., Shoichet, B.K., Krogan, N.J., 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583, 459–468. URL: https://doi.org/10.1038/s41586-020-2286-9, doi:10.1038/s41586-020-2286-9.

Guo, W., Kotsia, I., Patras, I., 2012. Tensor learning for regression. IEEE Transactions on Image Processing 21, 816–827. URL: https://doi.org/10.1109/tip.2011.2165291, doi:10.1109/tip.2011.2165291.

He, L., Lu, C.T., Ma, G., Wang, S., Shen, L., Yu, P.S., Ragin, A.B., 2017. Kernelized support tensor machines, in: Precup, D., Teh, Y.W. (Eds.), International Conference on Machine Learning, PMLR, International Convention Centre, Sydney, Australia. pp. 1442–1451. URL: http://proceedings.mlr.press/v70/he17a.html.

Liu, X., Guo, T., He, L., Yang, X., 2015. A low-rank approximation-based transductive support tensor machine for semisupervised classification. IEEE Transactions on Image Processing 24, 1825–1838.

Liu, Y., Shang, F., Fan, W., Cheng, J., Cheng, H., 2016. Generalized higher order orthogonal iteration for tensor learning and decomposition. IEEE Transactions on Neural Networks and Learning Systems 27, 2551–2563. URL: https://doi.org/10.1109/tnnls.2015.2496858, doi:10.1109/tnnls.2015.2496858.

Loh, W.Y., 2011. Variable selection for classification and regression in large p, small n problems, in: Probability Approximations and Beyond. Springer New York, pp. 135–159. URL: https://doi.org/10.1007/978-1-4614-1966-2_10, doi:10.1007/978-1-4614-1966-2_10.

Ng, K.L., Taguchi, Y.H., 2020. Identification of miRNA signatures for kidney renal clear cell carcinoma using the tensor-decomposition method. Scientific Reports 10. URL: https://doi.org/10.1038/s41598-020-71997-6, doi:10.1038/s41598-020-71997-6.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47–e47. URL: https://doi.org/10.1093/nar/gkv007, doi:10.1093/nar/gkv007, arXiv:https://academic.oup.com/nar/article-pdf/43/7/e47/7207289/gkv007.pdf.

Schölkopf, B., 2000. The kernel trick for distances, in: Proceedings of the 13th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA. p. 283–289.

Secchi, P., Stamm, A., Vantini, S., 2013. Inference for the mean of large p small n data: A finite-sample high-dimensional generaliza-

tion of hotelling's theorem. Electronic Journal of Statistics 7, 2005–2031. URL: https://doi.org/10.1214/13-ejs833, doi:10.1214/13-ejs833.

Signoretto, M., Lathauwer, L.D., Suykens, J.A., 2011. A kernel-based framework to tensorial data analysis. Neural Networks 24, 861–874. URL: https://doi.org/10.1016/j.neunet.2011.05.011, doi:10.1016/j.neunet.2011.05.011.

Signoretto, M., Olivetti, E., De Lathauwer, L., Suykens, J.A.K., 2012. Classification of multichannel signals with cumulant-based kernels. IEEE Transactions on Signal Processing 60, 2304–2314.

Taguchi, Y.H., 2017. Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases. BMC Medical Genomics 10. URL: https://doi.org/10.1186/s12920-017-0302-1, doi:10.1186/s12920-017-0302-1.

Taguchi, Y.H., 2020. Unsupervised Feature Extraction Applied to Bioinformatics. Springer International Publishing. URL: https://doi.org/10.1007/978-3-030-22456-1, doi:10.1007/978-3-030-22456-1.

Taguchi, Y.h., Turki, T., 2020. A new advanced in silico drug discovery method for novel coronavirus (sars-cov-2) with tensor decomposition-based unsupervised feature extraction. PLOS ONE 15, 1–16. URL: https://doi.org/10.1371/journal.pone.0238907, doi:10.1371/journal.pone.0238907.

Tao, D., Li, X., Wu, X., Hu, W., Maybank, S.J., 2007. Supervised tensor learning. Knowledge and Information Systems 13, 1–42. URL: https://doi.org/10.1007/s10115-006-0050-6, doi:10.1007/s10115-006-0050-6.

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences 98, 5116–5121. URL: https://www.pnas.org/content/98/9/5116, doi:10.1073/pnas.091062498, arXiv:https://www.pnas.org/content/98/9/5116.full.pdf.

Wang, X.D., Chen, R.C., Yan, F., Zeng, Z.Q., Hong, C.Q., 2019a. Fast adaptive k-means subspace clustering for high-dimensional data. IEEE Access 7, 42639–42651. URL: https://doi.org/10.1109/access.2019.2907043, doi:10.1109/access.2019.2907043.

Wang, X.D., Chen, R.C., Zeng, Z.Q., Hong, C.Q., Yan, F., 2019b. Robust dimension reduction for clustering with local adaptive learning. IEEE Transactions on Neural Networks and Learning Systems 30, 657–669. URL: https://doi.org/10.1109/tnnls.2018.2850823, doi:10.1109/tnnls.2018.2850823.

Yan, S., Xu, D., Yang, Q., Zhang, L., Tang, X., Zhang, H., 2007. Multilinear discriminant analysis for face recognition. IEEE Transactions on Image Processing 16, 212–220.

Zhao, Q., Zhou, G., Adali, T., Zhang, L., Cichocki, A., 2013a. Kernelization of tensor-based models for multiway data analysis: Processing of multidimensional structured data. IEEE Signal Processing Magazine 30, 137–148.

Zhao, Q., Zhou, G., Adalı, T., Zhang, L., Cichocki, A., 2013b. Kernel-based tensor partial least squares for reconstruction of limb movements, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3577–3581.

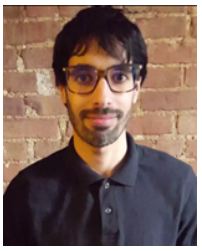
Acknowledgements

This work was supported by KAKENHI [grant numbers 19H05270, 20H04848, and 20K12067] to YT and Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah [grant number KEP-8-611-38] to TT.

Author contributions statement

YHT: Conceptualization, Formal analysis, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing. TT: Data curation, Writing – original draft, Writing – review & editing

Y-H. Taguchi received a B.S. degree in physics from the Tokyo Institute of Technology and a Ph.D. degree in physics from the Tokyo Institute of Technology. He is currently a full professor with the Department of Physics, Chuo University, Japan. His works have been published in leading journals such as Physical Review Letters, Bioinformatics, and Scientific Reports. His research interests include bioinformatics, machine-learning, and nonlinear physics. He is also an editorial board member of Frontiers in Genetics:RNA, PloS ONE, BMC Medical Genomics, Medicine (Lippincott Williams & Wilkins journal), BMC Research Notes, noncoding RNA (MDPI), and IPSJ Transaction on Bioinformatics.



Turki Turki received a B.S. degree in computer science from King AbdulAziz University, an M.S. degree in computer science from NYU.POLY, and a Ph.D. degree in computer science from the New Jersey Institute of Technology. He is currently an assistant professor with the Department of Computer Science, King Abdulaziz University, Saudi Arabia. His research interests include Artificial Intelligence (Tensor Learning, Machine Learning, Deep Learning) and Bioinformatics. His research studies have been published in journals such as Expert Systems with Applications, Frontiers in Genetics, Current Pharmaceutical Design, Computers in Biology and Medicine, and Genes. Dr. Turki has served on the program committees of several international conferences and is currently a review editor for Frontiers in Artificial Intelligence and Frontiers in Big Data. In addition, he is an editorial board member of Computers in Biology and Medicine and Sustainable Computing: Informatics and Systems.