Tensor decomposition based unsupervised feature extraction in big data analytics applied to prostate cancer multiomics data

Y-h. Taguchi
Department of Physics
Chuo University
Tokyo 112-8551, Japan
tag@granular.com

Turki Turki
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
tturki@kau.edu.sa

***Abstract*—Large *p* small *n* problem is a challenging problem in big data analytics. There are no de facto standard methods available to it. In this study, we propose a tensor decomposition (TD) based unsupervised feature extraction (FE) formalism applied to multiomics datasets, where the number of features is more than 100000 while the number of instances is as small as about 100. The proposed TD based unsupervised FE outperformed other conventional supervised feature selection methods, such as random forest, categorical regression (also known as analysis of variance, ANOVA), and penalized linear discriminant analysis when they are applied to not only multiomics datasets but also synthetic datasets. Genes selected by TD based unsupervised FE were biologically reliable. TD based unsupervised FE turned out to be not only the superior feature selection method but also the method that can select biologically reliable genes.**

***Index Terms*—tensor decomposition, unsupervised learning, prostate cancer, feature selection, big data analytics**

I. Introduction

The big data term often means that there are many number of instances as well as features [1], [2]. Typical big data are a few million images (photos) composed of several millions of pixels. The number of instances are as many as instances, often more than that. Most of recently developed machine learning methods including deep learning (DL) are aiming to tackle such kind of problems [1], [3]–[8]. On the other hand, although they are not as popular as these typical big data problems, there is another branch of "Big Data" problem which is known as "large *p* small *n*" problem or "high dimensional data analysis". For these cases, the number of instances is typically smaller than the number of features [9]–[11]. In this case, there can be many unique problems that do not exist for typical big data problems mentioned above. For example, in the discrimination tasks, since there are more features than instances, if we do not reduce the number of features, accuracy is always 100% which is nothing but overfitting. Alternatively, there can be the so-called "curse of dimensionality" problem that means short of instances covering large enough spaces. Because of these difficulties, compared with the popular big data problems, "large *p* small *n*" problem is hardly investigated in a comprehensive manner.

One might think that it is not an issue, since "large *p* small *n*" problem is relatively rare and is not so important. Nevertheless, dependent on the domains, "large *p* small *n*" problem would be inevitably typical. For example, in genomic science, the typical number of features is as many as or more than that of genes (i.e., about 10^4) while the number of instances is the number of patients, typically at most a few hundreds. The length of DNA sequence of human genome is about 3×10^9 while the human population is about 10^{10}. However, only very small part of it can be considered in individual study. Thus, in typical genomic studies, the ratio of the number of features to the number of samples can be as large as 10^2 or more than that.

In addition, there is yet another difficulty in genomic science: labels attributed to instances are often not numeric but categorical and even not ordered. This prevents us from applying most of popular methods (e.g., lasso [12]) to the genomic datasets. Thus, applicable methods are very limited.

On the other hand, Taguchi has proposed a very different method from the typical machine learning methods that are applicable to popular big data problems: tensor decomposition (TD) based unsupervised feature extraction (FE) [13]. In this method, smaller number of representative features, which are called singular value vectors, are generated with the liner combinations of original huge number of features, without considering labeling. These singular value vectors are attributed to both instances and original features. With investigating singular value vectors attributed to instances, we can identify which of them are useful (e.g., distinguising between instances associated with distinct labeling). Then we select singular value vectors associated with the selected singular value vectors attributed to instances among those attributed to the original features. Finally, the selected singular value vectors attributed to the original features are used to evaluate importance of the original features; since the singular value vectors are the linear combination of original features, their coefficients of linear combinations can be used for the

This study was supported by KAKENHI, 20K12067, 20H04848, and 19H05270

evaluation of importance of individual original features. The original features having larger absolute values of coefficients are selected. As a result, we can avoid many difficulties in "large *p* small *n*" problem, e.g., we can perform standard discriminant analysis when there are more numbers of features than instances.

In this paper, we propose a TD based unsupervised FE formalism, applied to synthetic data that imitate typical genomic datesets. The proposed method is first compared against three conventional methods applicable to "large *p* small *n*" problem: random forest (RF), penalized linear discriminant analysis (LDA) and categorical regression. Then, these four methods are also applied to a real dataset consisting of multiomics measurements of prostate cancers, a typical "large *p* small *n*" datasets. Results for synthetic and real datasets demonstrate the superiority of our proposed method when compared against the three conventional methods.

II. The Materials and Methods

A. Synthtic data

In the synthetic dataset, we assumed that there are 81 instances composed of three tissue subclasses (each class has three replicates) times three disease subclasses (each class also has three replicates). That is, 81 is composed of all pairs of one of nine tissue instances (= three tissues times three replicates) and one of nine disease instances (= three tissues times three replicates). Typically, three tissues are supposed to be, for example, heart, brain and skin while three diseases are supposed to be prostate cancer, diabetes and healthy controls; this means that we do not have any biological knowledge about which combinations are more likely associated with one another. The number of genes, which are features, is as many as 10^5; although so-called protein coding genes are at most a few tens thousands, more number of non-coding genes have come to be taken into account. Unlike synthetic data, in the real dataset to which the methods are applied in the later part of this paper, the number of features corresponds to the number of genomic regions with fixed DNA sequence length. Thus, it can be even much more than that of the synthetic data. Among these 10^5 features, only the first 100 features are supposed to be associated with distinction between diseases and tissues. The purpose of analysis is to identify these 100 features correctly based upon the available datasets in a fully data-driven approach. In order to emulate this situation, we introduced tensor $x_{ijk} \in \mathbb{R}^{10^4 \times 9 \times 9}$ as

$$x_{ijk} = \begin{cases} \lceil \frac{j}{3} \rceil \left(\lceil \frac{k}{3} \rceil + 3 \right) + \epsilon_{ijk}, & i \leq 100 \\ \epsilon_{ijk}, & i > 100 \end{cases} \tag{1}$$

where $\epsilon_{ijk} \sim \mathcal{N}(1, 0)$, $\mathcal{N}(\mu, \sigma)$ is Gaussian distribution with mean of μ and standard deviation of σ and $\lceil x \rceil$ is ceiling function that gives the smallest integer that is not less than x.

B. Real dataset

The real dataset consists of multiomics measurements of prostate cancer retrieved from gene expression omnibus (GEO) [14], using GEO ID GSE130408. The processed file named GSE130408_RAW.tar was downloaded and individual files whose names start with "GSM" were extracted from it. Individual files are composed of eight omics measurements: three transcription factor bindings, AR, FOXA1, HOXB13, four histone modification, H3K27AC, H3K27me3, H3K4me3, K4me2, and ATAC-seq. To each of them, two additional subclasses, tumor and normal tissue, were attributed. Since the number of biological replicates vary from subclass to subclass, we randomly selected the following number of ChIP-seq subclasses: two AR, four FOXA1, four HOXB13, ten H3K27AC, one H3K27me3, one H3K4me3, one K4me2, and one ATAC-seq. In total, 24 multiomics measurements groups were constructed. Each multiomics measurements group is composed of six instances, which are composed of two tissue subclasses (tumor and normal prostate), each of which has three biological replicates (Table I).

TABLE I
SUMARY OF REAL DATASET

Omics	number of multiomics	number of tissues	number of biological replicates
AR	2	2	3
FOXA1	4	2	3
HOXB13	4	2	3
H3K27AC	10	2	3
H3K27me3	1	2	3
H3K4me3	1	2	3
K4me2	1	2	3
ATAC	1	2	3
total	24	—	—

Values in each file were averaged over 25000 DNA sequence intervals, whose total number is 123817, since 3×10^9, which is the total DNA length of human genome, divided by 25000 is equal to 120000. The averaged values are treated as a dataset to be analysed further.

C. Categorical regression

Categorical regression, which is also known as ANOVA, is

$$x_{ijk} = \sum_{J=1}^{3} a_J \delta_{Jj} + \sum_{K=1}^{3} b_K \delta_{Kk} \tag{2}$$

where δ_{jJ} and δ_{kK} takes 1 only when j or k belongs to Jth tissue or Kth disease subclasses. a_J and b_K are regression coefficients (for synthetic data).

$$x_{ijkm} = \sum_{J=1}^{8} a_J \delta_{Jj} + \sum_{K=1}^{2} b_K \delta_{Kk} \tag{3}$$

where δ_{jJ} and δ_{kK} takes 1 only when j or k belongs to Jth multiomics measurement group or Kth tissue ($K = 1$:tumor, $K = 2$:normal prostate) subclasses. a_J and b_K are regression coefficients (for real data).

The computation was performed by `ls` function in R [15]. The computed P-values were adjusted by BH criterion and ith associated with adjusted P-values less than 0.01 were selected.

D. RF

RF was performed using `randomForest` function implemented in `randomForest` package [16] in R. Synthetic and real data were regarded as $3\times3=9$ and $8\times2=16$ subclasses, respectively, each of which corresponds to one of (I,J) pairs. Features included in OOB were selected with selecting features having non-zero importance given by `importance` function implemented in `randomForest` package.

E. PenalizeLDA

PenalizedLDA was performed using `PenalizedLDA.cv` function implemented in `PenalizedLDA` package [17] in R. Synthetic and real data were regarded as $3\times3=9$ and $8\times2=16$ subclasses, respectively, each of which corresponds to one of (I,J) pairs. λ is taken to be $0.01, 0.02, 0.03, 0.04$ for synthetic data. PenalizedLDA actually could not be performed for real dataset because of zero in-subclass standard deviations for some features (For more details, see [18]).

F. TD based unsupervised FE

Although it was fully described in the recently published book [13], here we propose a variant, which is outlined briefly. First, a dataset must be formated as a tensor. For synthetic data, $x_{ijk}\in\mathbb{R}^{10^5\times9\times9}$. For real data, $x_{ijkm}\in\mathbb{R}^{N\times24\times2\times3}$ represents the averaged value of ith interval of jth omics measurements group of kth tissue ($k=1$:tumor, $k=2$:normal prostae) of mth biological replicates ($1\le m\le3$). Here N is the total number of genomic region of 25000 DNA sequence length, 123817.

Then higher order singular value decomposition (HOSVD) [13] was applied to the tensor and we got

$$x_{ijk}=\sum_{\ell_1=1}^{3}\sum_{\ell_2=1}^{3}\sum_{\ell_3=1}^{10^5}G(\ell_1\ell_2\ell_3)u_{\ell_1 j}u_{\ell_2 k}u_{\ell_3 i} \tag{4}$$

for synthetic data. $G\in\mathbb{R}^{10^5\times9\times9}$ is core tensor and $u_{\ell_1 j}\in\mathbb{R}^{3\times3}, u_{\ell_2 k}\in\mathbb{R}^{3\times3}, u_{\ell_3 i}\in\mathbb{R}^{10^5\times10^5}$ are singular value matrices that are orthogonal matrices. On the other hand, for real dataset,

$$x_{ijkm}=\sum_{\ell_1=1}^{24}\sum_{\ell_2=1}^{2}\sum_{\ell_3=1}^{3}\sum_{\ell_4=1}^{N}G(\ell_1\ell_2\ell_3\ell_4)u_{\ell_1 j}u_{\ell_2 k}u_{\ell_3 m}u_{\ell_4 i} \tag{5}$$

$G\in\mathbb{R}^{N\times24\times2\times3}$ is a core tensor and $u_{\ell_1 j}\in\mathbb{R}^{24\times24}, u_{\ell_2 k}\in\mathbb{R}^{2\times2}, u_{\ell_3 m}\in\mathbb{R}^{3\times3}, u_{\ell_4 i}\in\mathbb{R}^{N\times N}$ are singular value matrices that are orthogonal matrices.

The first step is to check which singular value vectors attributed to instance subclasses ($u_{\ell_1 j}$ attributed to tissue subclasses and $u_{\ell k}$ attributed to disease subclasses for synthetic dataset, $u_{\ell_1 j}$ attributed to multiomics measurement groups and $u_{\ell_2 k}$ attributed to tissues for real dataset) represents distinction between instances of interests. After identifying biologically interesting singular value vectors attributed to instances ($u_{\ell_1 j}$ and $u_{\ell_2 k}$ for synthetic data and $u_{\ell_1 j}$, $u_{\ell_2 k}$ and $u_{\ell_3 m}$ for real data), we next try to find singular value vectors attributed to features ($u_{\ell_3 i}$ for synthetic data and $u_{\ell_4 i}$ for real data) that share core tensor G with larger absolute values with those identified singular value vectors attributed to instances. For example, suppose that ℓ'_1,ℓ'_2 are selected for synthetic data and ℓ'_1,ℓ'_2,ℓ'_3 are selected for real data. In this case we seek ℓ_3 that has $G(\ell'_1\ell'_2\ell_3)$ with the larger absolute values for synthetic data and ℓ_4 that has $G(\ell'_1\ell'_2\ell'_3\ell_4)$ with the larger absolute values for synthetic data. Then using selected $u_{\ell_3 i}$ and $u_{\ell_4 i}$, P-values are attributed to ith feature as

$$P_i=P_{\chi^2}\left[>\left(\frac{u_{\ell_3 i}}{\sigma_{\ell_3}}\right)^2\right] \tag{6}$$

for synthetic data and

$$P_i=P_{\chi^2}\left[>\left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}}\right)^2\right] \tag{7}$$

for real data. $P_{\chi^2}[>x]$ is the cumulative probability of χ^2 distribution that the argument is larger than x. σ_{ℓ_3} and σ_{ℓ_4} are the standard deviations. The computed P-values are corrected by BH criterion and features associated with adjusted P-values less than 0.01 were selected.

III. Results

Figure 1 shows the flowchart performed in this study.

A. Synthetic data

We have applied RF, PenalizedLDA, categrical regression and TD based unsupervised FE to the synthetic dataset. When RF was applied to synthetic dataset, the features having non-zero importance are as many as 7289 although the 100 features having subclass dependence were correctly selected. Thus, RF has the ability to select all of correct features. Since it selects too many false positives, it is ineffective. On the other hand, when PenalizedLDA was applied to synthetic dataset, the 100 features having subclass dependence were selected and no false positives are selected. Thus, PenalizedLDA could outperform RF. The only disadvantage is that we have to find the best λ, which is the weight coefficient of L1 norm. After trial and errors, we found that when λ is taken to be 0.02, 0.03, and 0.04, it achieved the perfect performance mentioned in the above. Nevertheless, if we select λ is 0.01, it selects more than 200 featured although the 100 features having subclass dependence are correctly selected. If we do not have pre-knowldge that the correct number is 100, it is impossible to tune the best parameter, since both $\lambda=0.01$ and $\lambda=0.04$ achieve the perfect prediction accuracy of discrimination of 81 instances into nine subclasses (In average, only one instance among 81 instances is wrongly classified). Thus, in this sense, PenaizedLDA cannot be regarded as the perfect method although it can achieve perfect performance if we can successfully tune the best parameter. Categorical regression is much better than PenalisedLDA, since it achieves the good performances regardless the threshold P-value (Table II). At most it has as small as 11 false positives even if the threshold P-values vary from 0.01 to 0.1. It is very contrast to PenaliazedLDA as the number of false positives varies from

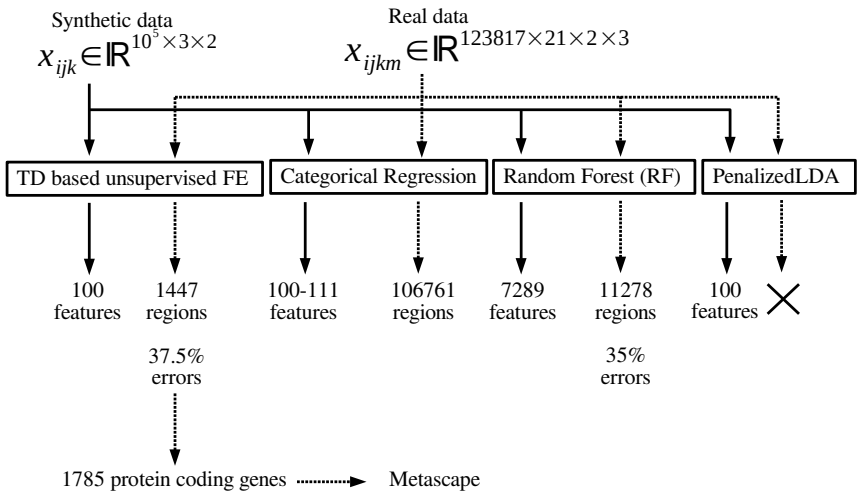

Fig. 1. Flowchart of analyses performed in this study

TABLE II
CONFUSION MATRICES WHEN CATEGORICAL REGRESSION IS APPLIED TO SYNTHETIC DATA. P IS THE THRESHOLD P-VALUES. FEATURED ARE SELECTED WHEN ADJUSTED P-VALUES LESS THAN THE THRESHOLD P-VALES ARE ASSOCIATED WITH THEM.

P	0.1		0.05		0.01		0.001	
	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$
not selected	99889	0	99895	0	99899	0	99900	0
selected	11	100	5	100	1	100	0	100

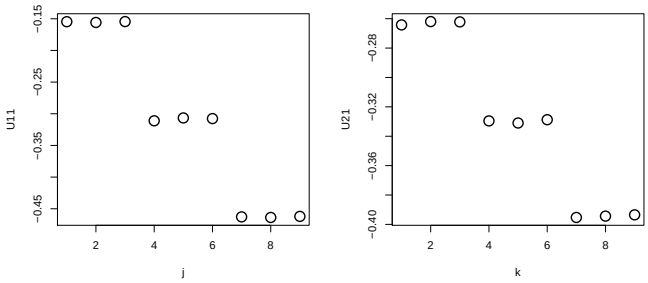

Fig. 2. u_{1j} and u_{2k} obtained when HOSVD applied to synthetic dataset.

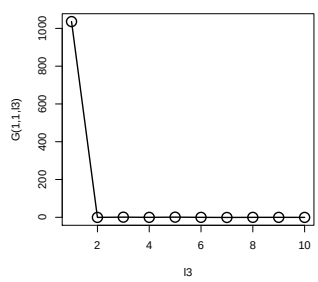

Fig. 3. $G(1, 1, \ell_3)$ obtained when HOSVD was applied to synthetic dataset.

0 to more than one hundred when λ slightly decreases from 0.02 to 0.01.

Nevertheless, TD based unsupervised FE turns out to even outperform categorical regression. At first, after applying HOSVD to synthetic dataset, we noticed that u_{1j} and u_{2k} are associated with distinction between three subclasses (Fig. 2). Thus we tried to find which ℓ_3 is associated with $G(1, 1, \ell_3)$ of the largest absolute value (Fig. 3). Then it was obvious that $G(1, 1, 1)$ has the largest absolute values. Equation (6) with $\ell_3 = 1$ was used to attribute P_i to ith feature. P_i was corrected by BH criterion and featured associated with adjusted P-values less than threshold values are selected (Table III). As a result, TD based unsupervised FE achieved perfect performances regardless the threshold P-values. Thus, it is obvious that TD based unsupervised FE is the best method among the four methods tested toward synthetic datasets.

Although the above results are enough to insist that TD based unsupervised FE is the best, it has yet another advantage: cpu time (Table IV). TD based unsupervised FE is the fastest method among the four methods tested toward the synthetic data. It can finish within a few seconds while others take more ranging from a few tens seconds for categorical regression to a few tens minutes for RF. In this sense, TD based unsupervised FE is the fastest, the most accurate and the most robust method among the four methods tested toward the synthetic dataset.

B. Real data

Next, we apply TD based unsupervised FE to a real dataset, since it was the best method among the four methods tested toward synthetic data in the previous subsection. After applying

TABLE III

CONFUSION MATRICES WHEN TD BASED UNSUPERVISED FE IS APPLIED TO SYNTHETIC DATA. P IS THE THRESHOLD P-VALUES. FEATURED ARE SELECTED WHEN ADJUSTED P-VALUES LESS THAN THE THRESHOLD P-VALES ARE ASSOCIATED WITH THEM.

P	0.1		0.05		0.01		0.001	
	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$
not selected	99900	0	99900	0	99900	0	99900	0
selected	0	100	0	100	0	100	0	100

TABLE IV

CPU TIMES OF FOUR METHODS REQUIRED FOR SYNTHETIC DATASET AND REAL DATASET. CPU TIME FOR PENALIZED LDA IS FOR TESTING 4 λ VALUES WITH ONE TO EIGHT DISCRIMINANT FUNCTIONS. CPU TIME FOR INDIVIDUAL RUN IS 18 SEC.

	methods	TD based unsupervised FE	categorical regression	random forest	penalized LDA
cpu time [sec]	synthetic data	6	66	142	575 (18)
	real data	20	87	321	—

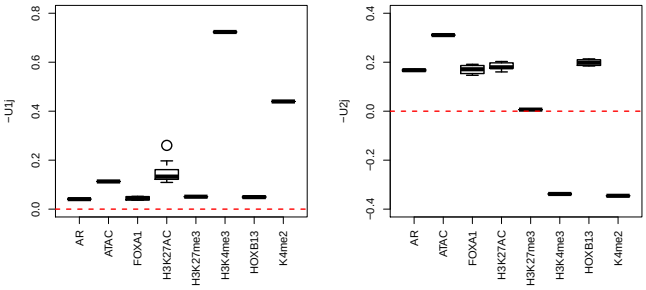

Fig. 4. Boxplot of u_{1j} and u_{2j} when HOSVD was applied to real dataset.

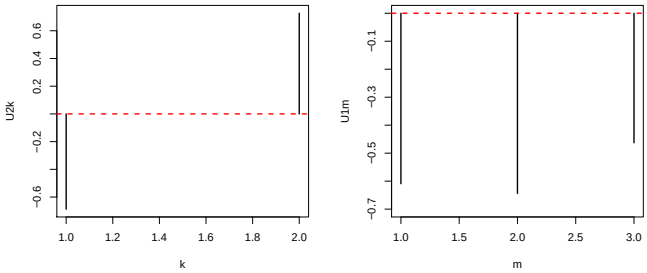

Fig. 5. u_{2k} and u_{1m} when HOSVD was applied to real dataset.

HOSVD to x_{ijkm}, we realized that u_{1j} and u_{2j} have biological significant dependence upon mulitomics measurements groups (Fig. 4). In order to explain why u_{1j} and u_{2j} are biologically reliable, we need to introduce genomic science briefly as follows:

Genome, which is a sequence of DNA, stores the information to be translated into protein. Since all of cells store the same genome, distinct functions and structures of individual cells are controlled by which part of genome are translated into proteins through transcription to RNA. This process is controlled by the various proteins that bind to genomes. In this measurements, the amount of three such proteins, AR, FOXA1, HOXB13, are measured. Yet another factors that control this process are the small molecules that bind to histone. Histone is a core protein around which DNA wraps and how tightly DNA wraps around histone is the factor that controls the process from DNA to RNA. H3K27AC, H3K27me3, H3K4me3, K4me2, are the measure which small molecules bind to which part of histone. Finally ATAC-seq measures how "OPEN" DNA is. As DNA becomes more open, it can produce more RNAs. Thus, the amount of proteins that bind to DNA, that of small molecules that bind to histone, and ATAC-seq are the measurement on how much genomes are transcribed to RNAs. The difficulty of multiomics analysis is to identify what kind of combinations of protein binding to DNA and small molecules binding to histone can enhance openness of DNA, which is assumed to be observed by ATAC-seq. Since we do not know what combination is likely observed, supervised approach is not easy to be employed.

Now we can explain why u_{1j} and u_{2j} are biologically reliable. H3K4me3 and H3K27ac are known to be active/open chromatin marks while H3K27me3 is known to be inactive marks [19]. u_{1j} has relatively larger values for H3K4me3 and H3K27ac while smaller values for H3K27me3 with having larger values for ATAC-seq, which measures the amount of openness of genome. Thus, the dependence of u_{1j} upon H3K4me3, H3K27ac, H3K27me3, and ATAC-seq are reliable. H3K4me2 [20] is also known to activate gene transcription in tissue specific ways. Overall dependence of u_{1j} upon multiomics measurements groups are reasonable. On the other hand, three proteins, AR [21], FOXA1 [22], and HOXB13 [23] are also known to mediate prostate cancer progression, their enrichment was coincident by u_{2j}. We decided that u_{1j} and u_{2j} successfully captured the biological aspects of multomics measurments.

On the other hand, u_{2k} represents distinction between tumor ($k = 1$) and normal prostate ($k = 2$) while u_{1m} represents independence of biological replicates (Fig. 5). Since we seek multiomics measurements that are distinct between tumor and normal prostate as well as common between biological replicates, u_{2j} and u_{1m} represent the property that we seek. Although we can select either $\ell_1 = 1$ or $\ell_1 = 2$ since both are biologically reliable, we select $\ell_1 = 1$ since it is the first (i.e., mathematically primary) factor. In order to find $u_{\ell_4 i}$ used for

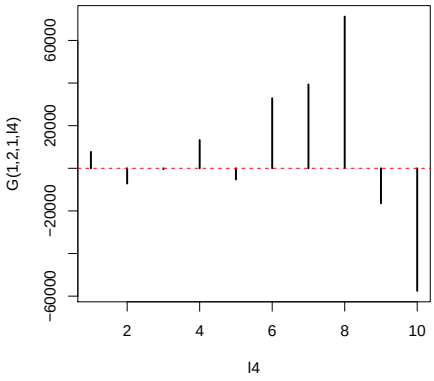

Fig. 6. $G(1, 2, 1, \ell_4)$ when HOSVD was applied to real dataset.

attributing P-values to genes i, we seek $G(1, 2, 1, \ell_4)$ having the largest absolute value (Fig. 6).

It is obvious that $G(1, 2, 1, 8)$ has the largest absolute value. Then, u_{8i} was employed to assign P-values to ith genomic regions using (7) with $\ell_4 = 8$. P_is were corrected with BH criterion and 1447 genomic regions associated with adjusted P_i less than 0.01 were selected among total 123817 genomic regions. This means that TD based unsupervised FE selected about as small as 1 % genomics regions among whole human genome.

Various biological evaluations of selected regions are possible. One of them is how many protein coding regions are included in these selected 1447 regions. Since genomic regions that include protein coding genes are less than a few percentages of human genome and total number of human protein coding genes is as many as 2×10^4, expected number of protein coding genes included in these selected 1447 genomic regions is at most a few hundreds. Nevertheless, if we count protein coding genes included in these regions, it is as many as 1785 genes, which is much larger than expected numbers. This indicates that TD based unsupervised FE can select genomic regions that include protein coding genes.

Another evaluation of these selected 1447 genomic regions is the biological properties of 1785 protein coding genes included in these 1447 genomic regions. If biologically reliable protein coding genes are selected, they should be enriched with the genes related to prostate cancer. In order to confirm this point, we uploaded these genes to Metascape [24] that evaluates enrichment of various biological terms with reducing redundancy. Figure 7 shows the results of DisGeNet [25] category of Metascape. It is obvious that prostatic neoplasms is not only top ranked but also outstandingly enriched. This suggests that TD based unsupervised FE could specifically select genes related to prostate cancer. Figure 8 shows the results of PaGenBase [26] category of Metascape. It is obvious that LNCAP is not only top ranked but also outstandingly enriched. LNCAP [27] is a model cell line of human prostatic

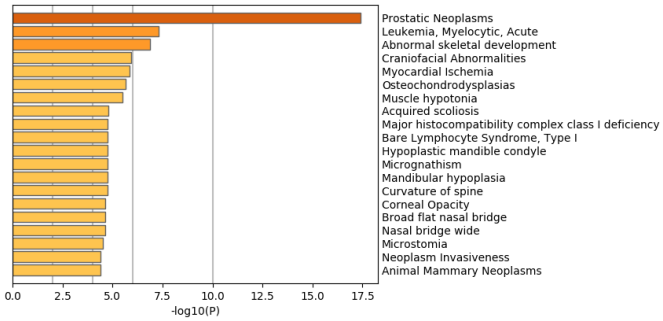

Fig. 7. Summary of enrichment of DisGeNet category of Metascape when 1785 protein coding genes included in 1447 genomic regions selected by TD based unsupervised FE are considered.

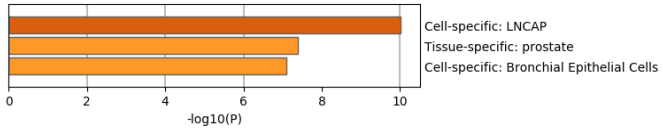

Fig. 8. Summary of enrichment of PaGenBase category of Metascape when 1785 protein coding genes included in 1447 genomic regions selected by TD based unsupervised FE are considered.

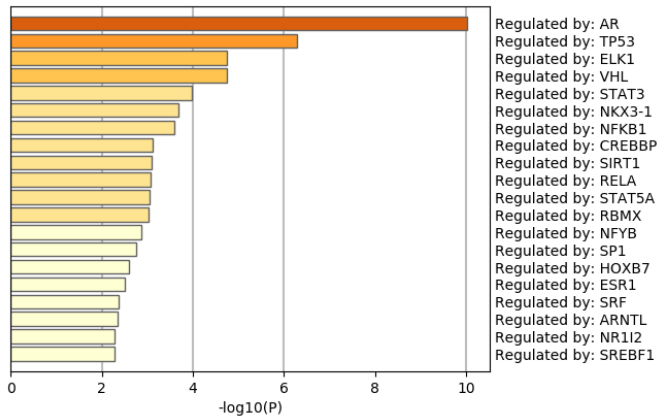

Fig. 9. Summary of enrichment of TRRUST category of Metascape when 1785 protein coding genes included in 1447 genomic regions selected by TD based unsupervised FE are considered.

carcinoma. This also suggests that TD based unsupervised FE could specifically select genes related to prostate cancer. Figure 9 shows the results of TRRUST [28] category of Metascape. It is obvious that AR is not only top ranked but also outstandingly enriched. As mentioned in the above, AR is a protein that binds to genome and mediates prostate cancer progression. This also suggests that TD based unsupervised FE could specifically select genes related to prostate cancer. Although there are more convincing results provided by Metascape, supplementary materials would include the full report provided by Metascape.

In the following, we would like to compare the performances of TD based unsupervised FE with the other three methods: RF, PenalizedLDA and categorical regression. At

TABLE V

CONFUSION MATRIX FOR RF APPLIED TO REAL DATASET. N:NORMAL PROSTATE, T:TUMOR. ERORR RATE IS 35 %.

	AR		ATAC		FOXA1		H3K27AC		H3K27me3		H3K4me3		HOXB13		K4me2		class
	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	error
AR_N	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1/6
AR_T	0	0	0	0	1	3	1	0	0	0	0	0	0	1	0	0	1
ATAC_N	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	1
ATAC_T	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	2/3
FOXA1_N	1	0	0	0	9	0	0	0	0	0	0	0	2	0	0	0	1/4
FOXA1_T	0	0	0	0	4	5	0	0	0	0	0	0	0	3	0	0	7/12
H3K27AC_N	0	0	0	0	0	0	26	4	0	0	0	0	0	0	0	0	2/15
H3K27AC_T	0	0	0	0	0	0	6	24	0	0	0	0	0	0	0	0	1/5
H3K27me3_N	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	2/3
H3K27me3_T	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	1/3
H3K4me3_N	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1
H3K4me3_T	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1
HOXB13_N	0	0	0	0	2	0	0	0	0	0	0	0	9	1	0	0	1/4
HOXB13_T	0	0	0	0	0	0	0	0	0	0	0	0	1	11	0	0	1/12
K4me2_N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1
K4me2_T	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	1

TABLE VI

CONFUSION MATRIX FOR LDA USING $u_{1j}u_{2k}u_{3m}$ GIVEN BY TD APPLIED TO 1447 GENOMIC REGIONS SELECTED BY TD BASED UNSUPERVISED FE APPLIED TO REAL DATASET. N:NORMAL PROSTATE, T:TUMOR. ERROR RATE IS 37.5 %.

	AR		ATAC		FOXA1		H3K27AC		H3K27me3		H3K4me3		HOXB13		K4me2		class
	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	error
AR_N	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0	0	1
AR_T	0	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0	1
ATAC_N	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	1
ATAC_T	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	1
FOXA1_N	0	0	0	0	9	0	1	0	0	0	0	0	2	0	0	0	1/4
FOXA1_T	0	0	0	0	0	9	0	1	0	0	0	0	0	2	0	0	1/4
H3K27AC_N	0	0	2	0	0	0	25	0	0	0	1	0	2	0	0	0	1/6
H3K27AC_T	0	0	0	2	0	0	0	25	0	0	0	1	0	2	0	0	1/6
H3K27me3_N	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
H3K27me3_T	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
H3K4me3_N	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	1/3
H3K4me3_T	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	1/3
HOXB13_N	0	0	0	0	4	0	2	0	0	0	0	0	6	0	0	0	1/2
HOXB13_T	0	0	0	0	0	4	0	2	0	0	0	0	0	6	0	0	1/2
K4me2_N	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	1
K4me2_T	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1

first, RF was applied to the real dataset that is classified into sixteen categorical classes composed of any pairs of eight mulitiomics measurements groups and two tissues (tumor and normal prostate). RF results in as many as 11278 genomic regions as having none zero importance. This number is almost eight time larger than the number of regions selected by TD based unsupervised FE, 1447. One of the evaluations of how good RF performed is the accuracy of discrimination, since RF is a supervised method. RF shold have tried to select the minimum features that can successfully discriminate between sixteen classes. The error rate given by RF was as small as 0.35 (Table V). This means that 65 % of 144 instances are correctly classified into 16 classes (see Table I, how 144 instances are classified into 16 subclasses). In order to validate discrimination performances of 1447 genomic regions selected by TD based unsupervised FE, we have done the following. u_{1j}, u_{2k}, u_{3k} were recomputed by HOSVD using only 1447 genomic regions selected by TD based unsupervised FE. Then 144 instances were discriminated using liner discriminant analysis with employing the product $u_{1j}u_{2k}u_{3k}$ as input vectors (Table VI with leave-one-out cross-validation). The error rate was as small as 37.5 %, which is similar to that of RF, 0.35. With considering the fact that TD based unsupervised FE has employed only one eighth genomic regions of those selected by RF, we concluded that TD based unsupervised FE could outperform RF.

Next, we applied PenalizedLDA to the real dataset. Nevertheless, we found that we could not apply it to the real dataset. In order to perform LDA, all of in-subclass variance must not be zero. This condition was not fulfilled for real data. Because of this, PenalizedLDA could not be employed to process real datasets.

Finally, we applied categorical regression to the real dataset with regarding it as categorical 16 subclasses. The number of features associated with adjusted P-values less than 0.01 was as many as 106761 among all 123817 features. This means that categorical regression selected almost all feature and is clearly ineffective. The reason why categorical regression failed in spite of that it was relatively successful (Table II) when applied to synthetic dataset is as follows. In synthetic

dataset, there was only one dependence upon subclasses. Selecting features having sublcass dependence automatically gives us targeted features. In real dataset, it cannot stand as it is. There are many variations of subclass dependence. In TD based unsupervised FE, by accessing dependence of singular value vectors upon subclasses (see Fig. 4), we can specify what kind of dependence upon subclassed should be considered. This allows us to identify restricted number of biologically reliable genomic regions. Categorical regression cannot distinguish between various subclass dependence and only could identify genomic regions associated with any kind of subclass dependence. This results in the identification of almost all genomic regions since most of them might be associated with some kind of subclass dependence. As a result, none of three methods (RF, PenalizedLDA and categorical regression) could compete with TD based unsupervised FE when they are applied to real dataset.

In additon to this, the cpu time required is also the shortest compared to the other three methods when applied to the real dataset (Table IV). From this point view, TD based unsupervised FE outperformed other methods.

SUPPLEMENTARY DATA

Supplementary data set is available at http://dx.doi.org/10.21227/rdmb-jm40.

REFERENCES

[1] A. N. Richter and T. M. Khoshgoftaar, "Efficient learning from big data for cancer risk modeling: a case study with melanoma," *Computers in biology and medicine*, vol. 110, pp. 29–39, 2019.

[2] M. G. Awan, T. Eslami, and F. Saeed, "GPU-daemon: GPU algorithm design, data management & optimization template for array based big omics data," *Computers in biology and medicine*, vol. 101, pp. 163–173, 2018.

[3] M. Nashaat, A. Ghosh, J. Miller, S. Quader, C. Marston, and J.-F. Puget, "Hybridization of active learning and data programming for labeling large industrial datasets," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 46–55.

[4] R. Shah, S. Zhang, Y. Lin, and P. Wu, "xSVM: Scalable distributed kernel support vector machine training," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 155–164.

[5] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

[6] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Computers in Human Behavior*, vol. 93, pp. 309–317, 2019.

[7] K. Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet Oncology*, vol. 20, no. 5, pp. e262–e273, 2019.

[8] T. Santosh, D. Ramesh, and D. Reddy, "LSTM based prediction of malaria abundances using big data," *Computers in Biology and Medicine*, p. 103859, 2020.

[9] J. Ge, X. Li, H. Jiang, H. Liu, T. Zhang, M. Wang, and T. Zhao, "Picasso: A sparse learning library for high dimensional data analysis in r and python." *J. Mach. Learn. Res.*, vol. 20, pp. 44–1, 2019.

[10] F. Wen, L. Chu, R. Ying, and P. Liu, "Fast and positive definite estimation of large covariance matrix for high-dimensional data analysis," *IEEE Transactions on Big Data*, 2019.

[11] S. Yang, J. Wen, X. Zhan, and D. Kifer, "ET-lasso: a new efficient tuning of lasso-type regularization for high-dimensional data," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 607–616.

[12] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[13] Y. h. Taguchi, *Unsupervised Feature Extraction Applied to Bioinformatics: A PCA and TD Based Approach*. Springer international, 2019.

[14] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 01 2002. [Online]. Available: https://doi.org/10.1093/nar/30.1.207

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[16] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[17] D. Witten, *penalizedLDA: Penalized Classification using Fisher's Linear Discriminant*, 2015, R package version 1.1. [Online]. Available: https://CRAN.R-project.org/package=penalizedLDA

[18] D. M. Witten and R. Tibshirani, "Penalized classification using fisher's linear discriminant," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 5, pp. 753–772, 2011.

[19] A. A. Igolkina, A. Zinkevich, K. O. Karandasheva, A. A. Popov, M. V. Selifanova, D. Nikolaeva, V. Tkachev, D. Penzar, D. M. Nikitin, and A. Buzdin, "H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 histone tags suggest distinct regulatory evolution of open and condensed chromatin landmarks," *Cells*, vol. 8, no. 9, 2019. [Online]. Available: https://www.mdpi.com/2073-4409/8/9/1034

[20] A. Pekowska, T. Benoukraf, P. Ferrier, and S. Spicuglia, "A unique H3K4me2 profile marks tissue-specific gene regulation," *Genome Research*, vol. 20, no. 11, pp. 1493–1502, 2010. [Online]. Available: http://genome.cshlp.org/content/20/11/1493.abstract

[21] K. Fujita and N. Nonomura, "Role of androgen receptor in prostate cancer: A review," *The World Journal of Men's Health*, vol. 37, no. 3, p. 288, 2019. [Online]. Available: https://doi.org/10.5534/wjmh.180040

[22] J. Gerhardt, M. Montani, P. Wild, M. Beer, F. Huber, T. Hermanns, M. Müntener, and G. Kristiansen, "FOXA1 promotes tumor progression in prostate cancer and represents a novel hallmark of castration-resistant prostate cancer," *The American Journal of Pathology*, vol. 180, no. 2, pp. 848–861, Feb. 2012. [Online]. Available: https://doi.org/10.1016/j.ajpath.2011.10.021

[23] H. I. Navarro and A. S. Goldstein, "HoxB13 mediates AR-V7 activity in prostate cancer," *Proceedings of the National Academy of Sciences*, vol. 115, no. 26, pp. 6528–6529, 2018. [Online]. Available: https://www.pnas.org/content/115/26/6528

[24] Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda, "Metascape provides a biologist-oriented resource for the analysis of systems-level datasets," *Nature Communications*, vol. 10, no. 1, Apr. 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-09234-6

[25] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, vol. 48, no. D1, pp. D845–D855, 11 2019. [Online]. Available: https://doi.org/10.1093/nar/gkz1021

[26] J.-B. Pan, S.-C. Hu, D. Shi, M.-C. Cai, Y.-B. Li, Q. Zou, and Z.-L. Ji, "PaGenBase: A pattern gene database for the global and dynamic understanding of gene function," *PLOS ONE*, vol. 8, no. 12, pp. 1–9, 12 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0080747

[27] J. S. Horoszewicz, S. S. Leong, E. Kawinski, J. P. Karr, H. Rosenthal, T. M. Chu, E. A. Mirand, and G. P. Murphy, "LNCaP model of human prostatic carcinoma," *Cancer Research*, vol. 43, no. 4, pp. 1809–1818, 1983. [Online]. Available: https://cancerres.aacrjournals.org/content/43/4/1809

[28] H. Han, H. Shim, D. Shin, J. E. Shim, Y. Ko, J. Shin, H. Kim, A. Cho, E. Kim, T. Lee, H. Kim, K. Kim, S. Yang, D. Bae, A. Yun, S. Kim, C. Y. Kim, H. J. Cho, B. Kang, S. Shin, and I. Lee, "TRRUST: a reference database of human transcriptional regulatory interactions," *Scientific Reports*, vol. 5, no. 1, Jun. 2015. [Online]. Available: https://doi.org/10.1038/srep11432