Advanced tensor decomposition-based integrated analysis of protein-protein interaction with cancer gene expression can improve coincidence with clinical labels

1st Y-h. Taguchi
Departemtn of Physics
Chuo Univeristy
Tokyo, Japan
tag@granular.com, ORCID: 0000-0003-0867-8986

2nd Turki Turki
Department of Computer Sciences
King Abdulaziz University,
Jeddah, Saudi Arabia
tturki@kau.edu.sa, ORCID: 0000-0003-3583-9722

***Abstract*—When cancer gene expression profiles obtained from The Cancer Genome Atlas (TCGA) are integrated with protein-protein interaction (PPI) information using tensor decomposition (TD), the coincidence between clinical labeling and latent variables provided by singular value decomposition (SVD) can increase, although PPI does not obviously contain information about clinical class outcomes. This suggests that TD has the power to extract hidden structure in PPI, which has biological meaning that can help gene expression better coincide with clinical information. From a practical standpoint, integrated analysis can also improve discrimination performance using class labels to a degree.**

***Index Terms*—tensor decomposition, clinical information, singular value decomposition, advanced integrated analysis**

I. Introduction

Network structure representation (NSR), a technique commonly used in bioinformatics, is known to be useful for integrating hidden structure in massive data sets. Furthermore, numerous studies [1]–[3] in this field are described as "network-based." Nevertheless, in these network-based approaches, data sets are often provided in a matrix format that must be transformed into NSR. There are various ways to convert a matrix to an NSR, which results in arbitrariness. Conversely, if we can convert a network structure into matrix format, this might be more convenient since more data sets are provided as matrices than as NSRs. In this paper, we convert network structures represented as matrices into a form that can be easily integrated with matrix data, such a gene expression profile.

II. Materials and Methods

Figure 1 presents the comparisons between the conventional network-based approach and the tensor-based approach used in this study.

This work was supported by KAKENHI (Grant Numbers 20K12067) to Y.-H.T.

A. Matrix representation of network

In this paper, we employed the simplest NSR [4], where $n_{ii'} \in \mathbb{R}^{N\times N}$ is 1 only when the node i and i' are connected with each other and otherwise is zero.

B. Integrated analysis of matrices and networks with tensor

To convert an NSR into a matrix starting with $n_{ii'}$, we applied singular value decomposition (SVD) to $n_{ii'}$ as

$$n_{ii'} = \sum_{\ell} \lambda_\ell u_{\ell i} u_{\ell i'} \tag{1}$$

where λ_ℓ is a singular value and $u_{\ell i} \in \mathbb{R}^{N\times N}$ are the singular value matrix and the orthogonal matrix.

On the other hand, we apply SVD to $x_{ij} \in \mathbb{R}^{N\times M}$, which represents the gene expression of the ith gene of the jth sample as

$$x_{ij} = \sum_{\ell} \lambda'_\ell u'_{\ell i} v_{\ell j} \tag{2}$$

where $u'_{\ell i} \in \mathbb{R}^{M\times N}$ and $v_{\ell j} \in \mathbb{R}^{M\times M}$ (here we assume $M < N$). Then we generate the tensor $x_{ijk} \in \mathbb{R}^{N\times L\times 2}$ using the first $L(< M)$ SVVs as:

$$x_{i\ell k} = \begin{cases} u_{\ell i}, & k = 1 \\ u'_{\ell i}, & k = 2 \end{cases} \tag{3}$$

to which HOSVD is then applied, giving:

$$x_{i\ell k} = \sum_{\ell_1}\sum_{\ell_2}\sum_{\ell_3} G(\ell_1\ell_2\ell_3)\tilde{u}_{\ell_1 i}\tilde{u}_{\ell_2 \ell}\tilde{u}_{\ell_3 k} \tag{4}$$

where $G \in \mathbb{R}^{N\times L\times 2}$ is a core tensor and $\tilde{u}_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $\tilde{u}_{\ell_2 \ell} \in \mathbb{R}^{L\times L}$, and $\tilde{u}_{\ell_3 k} \in \mathbb{R}^{2\times 2}$ are singular value matrices and are orthogonal matrices. Throughout this article, $L = 10$. Next, we project x_{ij} onto $\tilde{u}_{\ell_1 i}$ by

$$\tilde{v}_{\ell_1 j} = \sum_{i} \tilde{u}_{\ell_1 i} x_{ij} \tag{5}$$

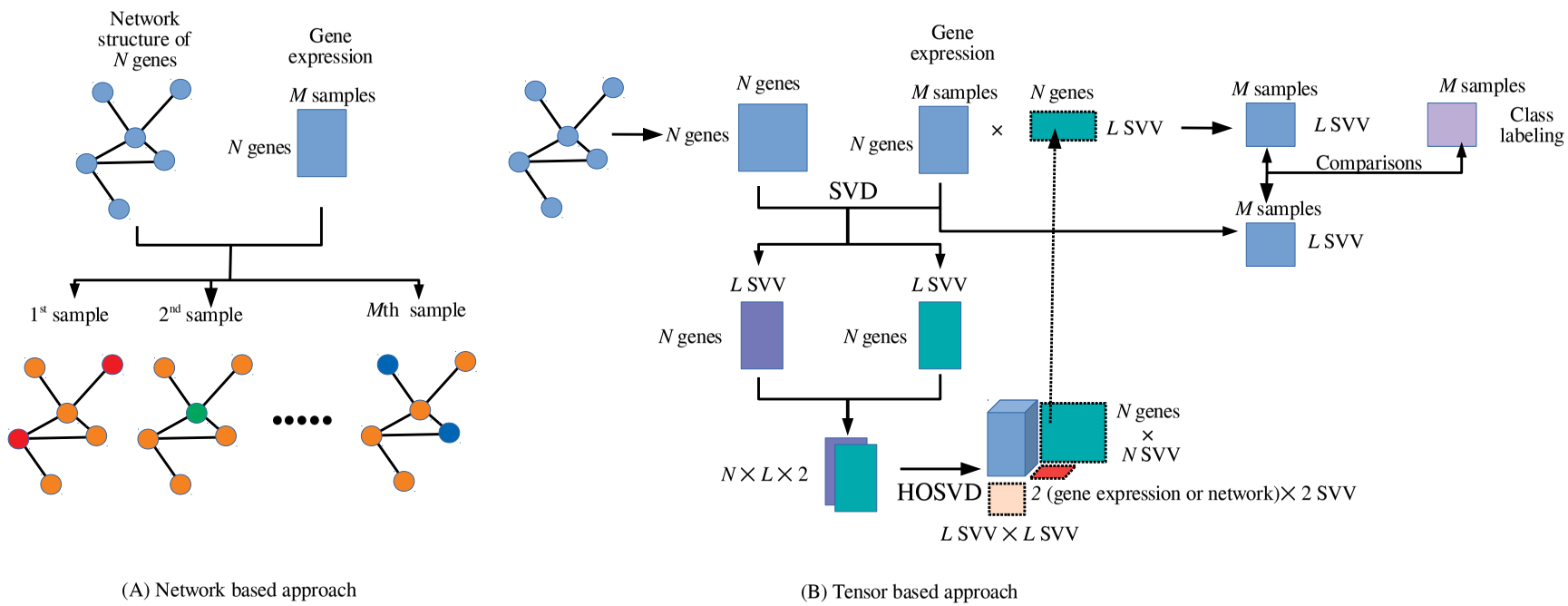

Fig. 1. Comparisons between (A) a network-based approach and (B) a tensor-based approach. Suppose that we have gene expression profiles and a network structure among genes. In the network-based approach, gene expression is represented by a network structure of genes. In the tensor-based approach, the network structure is converted to a matrix representation. SVD is applied to the matrix representation of the gene expression profiles and network structure. The obtained singular value vectors (SVVs) are bundled into a tensor to which HOSVD is applied. Gene expression profiles are projected to SVV and given by HOSVD, thus giving SVVs that represent sample dependence. The SVVs given by SVD and HOSVD are compared with class labels for validation.

C. Comparisons of coincidence with class labels between $v_{\ell j}$ and $\tilde{v}_{\ell_1 j}$

Categorical regression is performed for $v_{\ell j}$ and $\tilde{v}_{\ell_1 j}$ as

$$v_{\ell j} = a_\ell + \sum_s b_{\ell s}\delta_{js} \quad (6)$$

$$\tilde{v}_{\ell_1 j} = a'_{\ell_1} + \sum_s b'_{\ell_1 s}\delta_{js} \quad (7)$$

where δ_{js} is 1 only when j belongs to the sth class and otherwise is 0, and $a_\ell, b_{\ell s}, a'_{\ell_1}, b'_{\ell_1 s}$ are regression coefficients. P-values are computed by the lm function in R. The obtained P-values are corrected by the BH criterion [5] using the p.adjust function in R with the "BH" option.

If there are m data sets, then there are mL P-values, $P_h \in \mathbb{R}^{mL}$ which are ranked in ascending order (i.e., if $h > h'$ then $P_h > P_{h'}$). P_hs computed towards $v_{\ell j}$ and $\tilde{v}_{\ell_1 j}$ are compared with a paired t test and paired Wilcoxon test with the alternative hypothesis that the P_h found via SVD is greater than that given by HOSVD.

D. PPI data set

PPI data set, PP-Pathways_ppi.csv.gz, was retrieved from the human protein-protein interaction network [6], which includes 342,353 pairs for 21,557 proteins. After excluding self-pairs (i.e., self-dimers), this data was formatted as a $n_{ii'} \in \mathbb{R}^{N\times N}$ where $N = 21557$. Then there are 16,774 is which are common with the is in the TCGA gene expression profiles (see below). Since these 342,353 pairs represent only one of n_{ij} and n_{ji}, (i.e., pairs whose order is reversed are not included), when $n_{ij} \neq n_{ji}$, $n_{ij} = n_{ji} = 1$ is required.

E. Gene expression profile

TCGA gene expression profiles are used as x_{ij}. The RTCGA data set [7] was used for this purpose. RTCGA.rnaseq [8] was used as a gene expression profile. It includes 27 cancer data sets with various sample sizes (js) ranging from a few tens to a few hundred as well as 20,532 genes (is) whose expression profiles are available. The cancers considered are ACC, BLCA, BRCA, CESC, COAD, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, UCEC, and UCS. The class labels considered, retrieved from RTCGA.clinical [9], are patient.vital_status, patient.stage_event.pathologic_stage, patient.stage_event.tnm_categories. pathologic_categories. pathologic_m, patient.stage_event.tnm_categories.pathologic_categories. pathologic_n, and patient.stage_event.tnm_categories. pathologic_categories. pathologic_t. All 27 data sets are associated with "patient.vital_status" labels, patient.stage_event.pathologic_stage and patient.stage_event.tnm_categories.pathologic_categories. pathologic_m are associated with only 18 data sets, while patient.stage_event.tnm_categories. pathologic_categories.pathologic_t and patient.stage_event.tnm_categories. pathologic_categories.pathologic_n are associated with 20 data sets. (Table I).

16,774 is that are common with is in PPI (see above) are used.

F. Discriminant analysis

Linear discriminant analysis (LDA) for "patient.vital_status" was performed using the lda function in R. Leave-one-out cross-validation was employed with the setting "CV=T". By setting "prior = rep(1/2,2)", the two classes ("Dead'" and "Alive") are equally weighted, regardless of the number of samples within individual classes. LDA was performed using $v_{\ell j}$ of $\tilde{v}_{\ell_1 j}$ associated with adjusted P-values less than 0.05. When no SVVs were associated with adjusted P-values less than 0.05, LDA was not performed. AUC is computed by the colAUC function in the caTools package in R.

III. Results

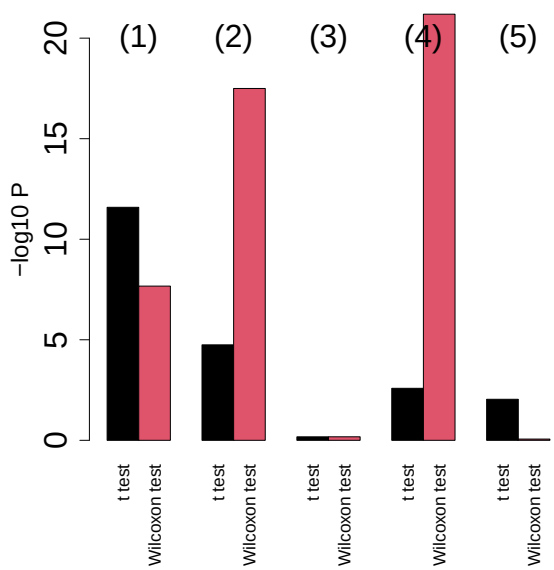

Fig. 2. Barplot of P-values computed by a t test and Wilcoxon test to evaluate the difference in P_h between SVD and HOSVD. (1) "patient.vital_status", (2) "patient.stage_event.pathologic_stage", (3) "patient.stage_event.tnm_categories.pathologic_categories.pathologic_m" (4) "patient.stage_event.tnm_categories.pathologic_categories.pathologic_t" (5) "patient.stage_event.tnm_categories.pathologic_categories.pathologic_n"

A. patient.vital_status

At first, we considered "patient.vital_status", which has two levels, "dead" and "alive." Figure 3 shows the scatter plot of P_h calculated from $v_{\ell j}$ and $\tilde{v}_{\ell_1 j}$, respectively. First, the values of P_h computed from $\tilde{v}_{\ell_1 j}$ are substantially smaller than those computed from $v_{\ell j}$. The P-values computed by

TABLE I

AVAILABILITY OF CLASS LABELS FOR CANCER DATA SETS. (1) "PATIENT.VITAL_STATUS", (2) "PATIENT.STAGE_EVENT.PATHOLOGIC_STAGE", (3) "PATIENT.STAGE_EVENT.TNM_CATEGORIES.PATHOLOGIC_CATEGORIES. PATHOLOGIC_M" (4) "PATIENT.STAGE_EVENT.TNM_CATEGORIES. PATHOLOGIC_CATEGORIES.PATHOLOGIC_T" (5) "PATIENT.STAGE_EVENT.TNM_CATEGORIES. PATHOLOGIC_CATEGORIES.PATHOLOGIC_N" (6) AUC FOR "PATIENT.VITAL_STATUS".

	ACC	BLCA	BRCA	CESC	COAD	ESCA	GBM	HNSC	KICH	KIRC	KIRP	LGG	LIHC	LUAD
(1)	○	○	○	○	○	○	○	○	○	○	○	○	○	○
(2)	○	○	○	×	○	○	×	○	○	○	○	×	○	○
(3)	×	○	○	○	○	○	×	○	○	○	○	×	○	○
(4)	○	○	○	○	○	○	×	○	○	○	○	×	○	○
(5)	○	○	○	○	○	○	×	○	○	○	○	×	○	○
(6)	×	○	○	×	×	×	×	○	×	○	×	○	○	○
	LUSC	OV	PAAD	PCPG	PRAD	READ	SARC	SKCM	STAD	TGCT	THCA	UCEC	UCS	Total ($= m$)
(1)	○	○	○	○	○	○	○	○	○	○	○	○	○	27
(2)	○	×	○	×	×	○	×	○	○	○	○	×	×	18
(3)	○	×	○	×	×	○	×	○	○	○	○	×	×	18
(4)	○	×	○	×	○	○	×	○	○	○	○	×	×	20
(5)	○	×	○	×	○	○	×	○	○	○	○	×	×	20
(6)	○	×	×	×	×	○	×	×	×	○	×	○	×	11

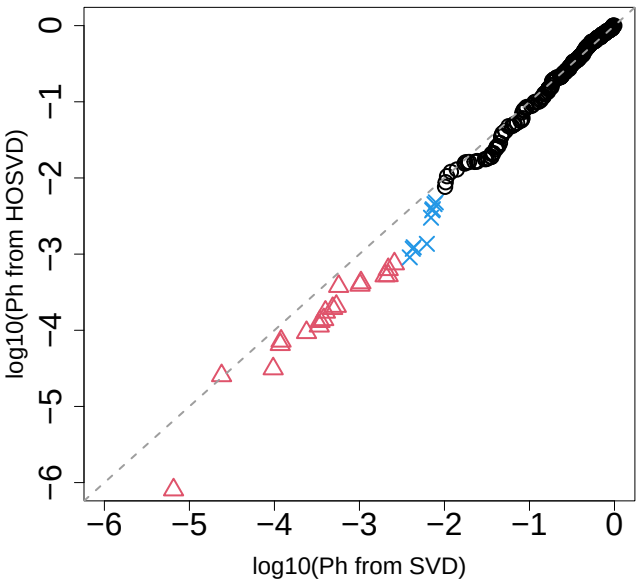

Fig. 3. Scatter plot (logarithmic scale) of P_h computed from $v_{\ell j}$ (horizontal axis) and $\tilde{v}_{\ell_1 j}$ (vertical axis) for patient.vital_status. Open red triangles are those associated with adjusted P-values less than 0.05 for both axes and blue crosses are those associated with adjusted P-values less than 0.05 for the vertical axis alone. Open black circles represent all other situations.

the paired t test and the paired Wilcoxon test applied to $\log_{10} P_h$ are 2.585×10^{-12} and 2.144×10^{-8}, respectively (Fig. 2). In addition, when the P_hs computed from $v_{\ell j}$ are not significant, the corresponding P_h computed from $\tilde{v}_{\ell_1 j}$ are not significant either (black open circles). This suggests that an integrated analysis with PPI can improve coincidence with class labels only when the P_h computed from SVD, i.e., without integration, is significant. This is natural, since PPIs are unlikely to have a coincidence with class labels as is.

B. *patient.stage_event.pathologic_stage*

Although for patient.vital_status labeling, an integrated analysis with PPI improved the coincidence of gene expression profiles with class labeling, one might wonder if this was accidental. In order to eliminate this possibility, we employed yet another class label, patient.stage_event.pathologic_stage, which reports cancer disease progression. Figure 4 shows the

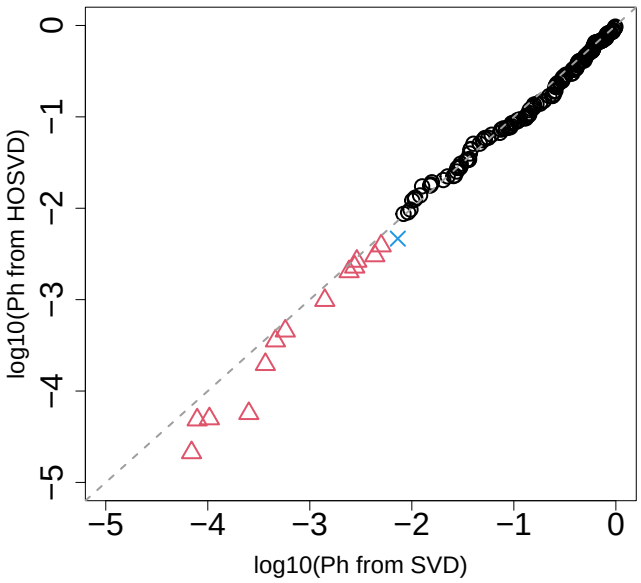

Fig. 4. Scatter plot in logarithmic scale of P_h computed from $v_{\ell j}$ (horizontal axis) and $\tilde{v}_{\ell_1 j}$ (vertical axis) for patient.stage_event.pathologic_stage. Open red triangles are those associated with adjusted P-values less than 0.05 for both axes and blue crosses are those associated with adjusted P-values less than 0.05 only for vertical axis. Open black circles represent all other situations.

scatter plot of P_h calculated from $v_{\ell j}$ and $\tilde{v}_{\ell_1 j}$, respectively. Again, the values of P_h computed from $\tilde{v}_{\ell_1 j}$ are substantially smaller than those from $v_{\ell j}$. The P-values computed by the

paired t test and the paired Wilcoxon test applied to $\log_{10} P_h$ are 1.791×10^{-5} and 3.178×10^{-18}, respectively (Fig. 2). When the P_hs computed from $v_{\ell j}$ are not significant, the corresponding P_h computed from $\tilde{v}_{\ell_1 j}$ are not either (black open circles), as in the case of patient.stage_event.pathologic_stage. Thus, the improvement of coincidence with class labeling is unlikely to be accidental.

C. patient.stage_event.tnm_categories.pathologic_categories. pathologic_m

One might think that PPI always improves the coincidence of gene expression profiles with class labeling, but this is not the case. When we consider patient.stage_event.tnm_categories.pathologic_categories. pathologic_m (Fig. 5), the values of P_h computed from $\tilde{v}_{\ell_1 j}$

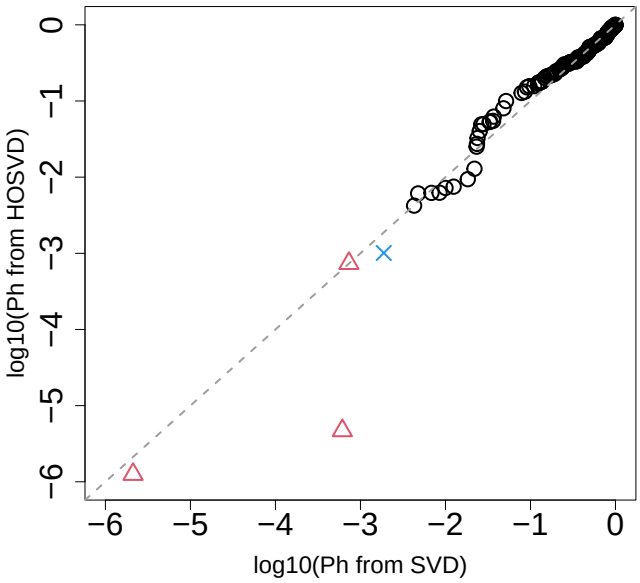

Fig. 5. Scatter plot (logarithmic scale) of P_h computed from $v_{\ell j}$ (horizontal axis) and $\tilde{v}_{\ell_1 j}$ (vertical axis) for patient.stage_event.tnm_categories. pathologic_categories.pathologic_m. Open red triangles are those associated with adjusted P-values less than 0.05 for both axes and blue crosses are those associated with adjusted P-values less than 0.05 for the vertical axis alone. Open black circles represent all other situations.

are not substantially smaller than those from $v_{\ell j}$; The P-values computed by paired t test and the paired Wilcoxon test applied to $\log_{10} P_h$ are 0.6714 and 0.6685, respectively (Fig. 2). The reason for this failure is obvious; there are only three $v_{\ell j}$s associated with adjusted P-values less than 0.05. As can be seen above, when SVD fails to attribute significant P-values to $v_{\ell j}$s, HOSVD also fails to attribute significant P-values to $\tilde{v}_{\ell_1 j}$s. Thus, for patient.stage_event.tnm_categories.pathologic_categories. pathologic_m, which cannot be associated with significant P-values by SVD, HOSVD fails as well.

D. patient.stage_event.tnm_categories.pathologic_categories. pathologic_t

When this labeling is considered, although the scatter plots (Fig. 6) between P_h computed by SVD (horizontal) and HOSVD (vertical) do not look substantially different, since the

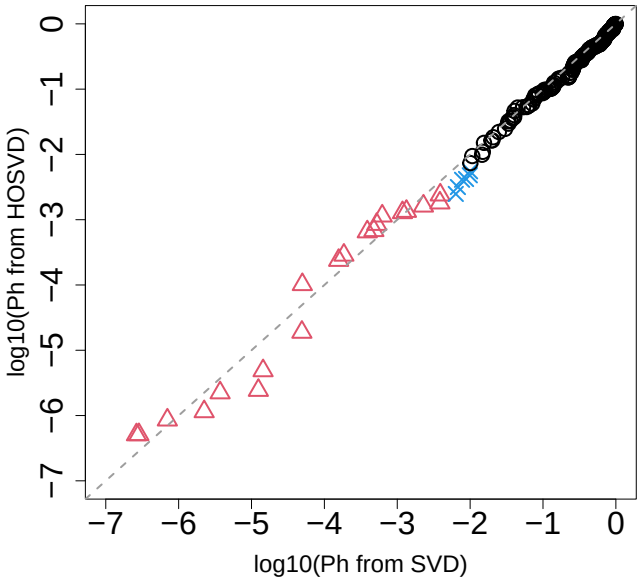

Fig. 6. Scatter plot (logarithmic scale) of P_h computed from $v_{\ell j}$ (horizontal axis) and $\tilde{v}_{\ell_1 j}$ (vertical axis) for patient.stage_event.tnm_categories. pathologic_categories.pathologic_t. Open red triangles are those associated with adjusted P-values less than 0.05 for both axes and blue crosses are those associated with adjusted P-values less than 0.05 for the vertical axis alone. Open black circles represent all other situations.

P-values computed by the paired t test and paired Wilcoxon test applied to $\log_{10} P_h$ are 2.591×10^{-3} and 6.430×10^{-22}, respectively (Fig. 2), integrated analysis with PPI surely improves the coincidence with class labels. This is possibly because the P-values computed from SVD are smaller than those in the previous three cases. In any case, when SVD can provide us a significant overlap with class labeling, HOSVD can do so as well.

E. patient.stage_event.tnm_categories.pathologic_categories. pathologic_n

When considering these class labels, although fewer $v_{\ell j}$ are associated with significant P-values, HOSVD can reduce the associated P-values substantially (Fig. 7). Reflecting this, the P-values computed by the paired t test and the paired Wilcoxon test applied to $\log_{10} P_h$ are 9.111×10^{-3} and 0.8667, respectively (Fig. 2); the small number of significant P-values prevents the latter test from detecting a significant difference $\log_{10} P_h$ between HOSVD and SVD. In any case, the t test can detect the presence of a significant difference.

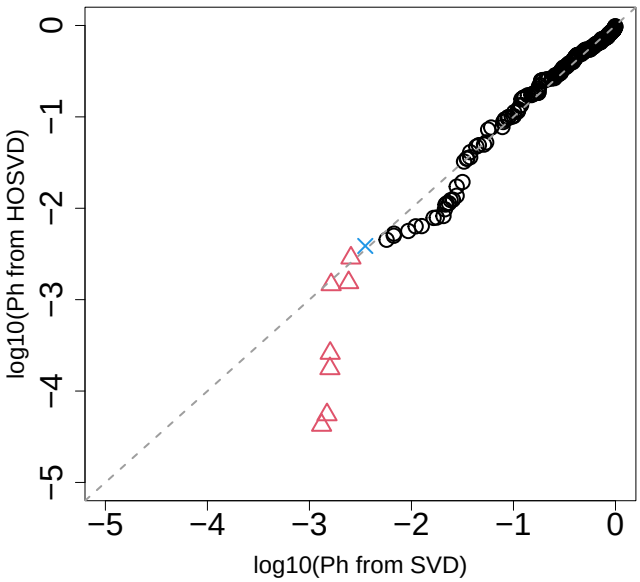

Fig. 7. Scatter plot (logarithmic scale) of P_h computed from $v_{\ell j}$ (horizontal axis) and $\tilde{v}_{\ell_1 j}$ (vertical axis) for patient.stage_event.tnm_categories. pathologic_categories.pathologic_t. Open red triangles are those associated with adjusted P-values less than 0.05 for both axes and blue crosses are those associated with adjusted P-values less than 0.05 for the vertical axis alone. Open black circles represent all other situations.

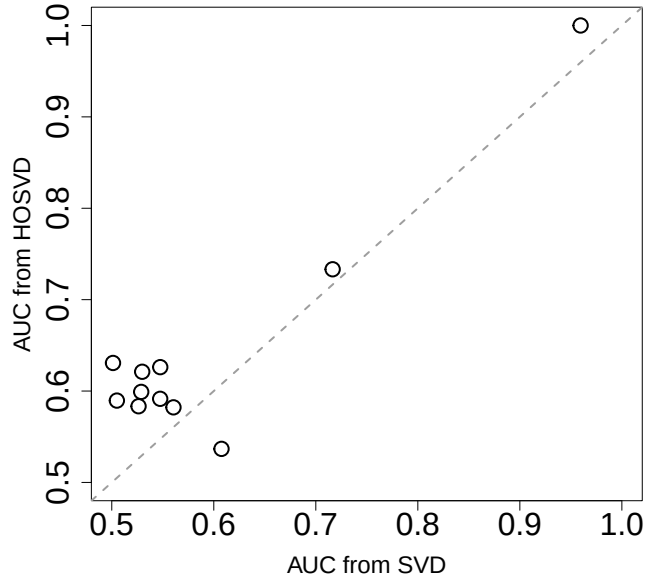

Fig. 8. Scatter plot of AUC for 11 cancer data sets (Table I) using the patient.vital_status label. SVD: horizontal, HOSVD: vertical. Ten of the 11 cancer data sets tested have a higher AUC for HOSVD.

F. Comparison of AUC for patient.vital_status

One might wonder whether the previous improvement is beneficial, since simply increasing the coincidence with the class labels might not be practically useful. In order to see if the improvement in the coincidence is effective, we consider a discrimination task using class labels. For this purpose, we consider the patient.vital_status label, since it has most cancer data (27) associated with labels within the five labels considered and it has only two classes ("Alive" and "Dead"), which allows us to compute AUC by which we can easily evaluate performance. Figure 8 shows the AUC computed for discrimination of 27 cancer data sets under the label of patient.vital_status. Only 11 cancer data sets (Table I) have SVV associated with significant P-values that coincide with the class labels. Although the AUC values themselves are not large and are mostly less than 0.6, the integrated analysis increases AUC; 10 out of 11 cancer data sets tested have a larger AUC for HOSVD. The P-values computed by the t test and Wilcoxon test with the alternative hypothesis that the AUC of HOSVD is larger than that by SVD are 5.744×10^{-2} and 3.417×10^{-3}, respectively. Although we cannot say that the P-values are very small, the result suggests that integrated analysis can improve the performance evaluated by AUC to some extent. Thus, integrated analysis with PPI is also practically useful.

IV. Discussion

In this study, we proposed an advanced tensor-based approach to integrate gene expression profiles with network structure. In contrast to the network-based approach (Fig. 1(A)) where individual gene expression profiles are directly embedded in the network structure, the tensor-based approach (Fig. 1(B)) can integrate network structure within the matrix format. It can automatically embed the network structure within gene expression profiles by converting the network structure to a low-dimensional-space representation.

As seen in Figs. 3 to 7, integrated analyses employing network structure can enhance the coincidence between latent variables and class labeling with information that seemingly is not included in the network structure. The SVVs derived by SVD from the representation of the PPI matrix are supposed to be vector representations of genes whose inner product is related to the similarity between genes expressed by the structure of the network. Since representation learning [10] enables us to convert diverse non-numerical information into a vector format, integration with tensors can be used for broader ranges of applications other than integration between PPI and gene expression. More studies are expected.

V. Conclusions

In this paper, we proposed an advanced tensor-based integrated approach and applied it to the integration of gene expression profiles and PPI. This can improve the coincidence between class labeling and latent variables, especially when latent variables derived without integration have significant coincidence with class labeling.

REFERENCES

[1] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, Dec. 2010. [Online]. Available: https://doi.org/10.1038/nrg2918

[2] F. Cheng, R. J. Desai, D. E. Handy, R. Wang, S. Schneeweiss, A.-L. Barabài, and J. Loscalzo, "Network-based approach to prediction and population-based validation of in silico drug repurposing," *Nature Communications*, vol. 9, no. 1, p. 2691, 2018.

[3] J. Ruan, A. K. Dean, and W. Zhang, "A general co-expression network-based approach to gene expression analysis: comparison and applications," *BMC Systems Biology*, vol. 4, no. 1, p. 8, 2010.

[4] M. Nakerekanti and V. Narasimha, "Analysis on malware issues in online social networking sites (sns)," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 335–338.

[5] Y.-H. Taguchi, *Unsupervised Feature Extraction Applied to Bioinformatics*. Springer International Publishing, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-22456-1

[6] (2018) Human protein-protein interaction network. [Online]. Available: https://snap.stanford.edu/biodata/datasets/10000/10000-PP-Pathways.html

[7] M. Kosinski, *RTCGA: The Cancer Genome Atlas Data Integration*, 2022, r package version 1.28.0. [Online]. Available: https://rtcga.github.io/RTCGA

[8] ——, *RTCGA.rnaseq: Rna-seq datasets from The Cancer Genome Atlas Project*, 2022, R package version 20151101.28.0.

[9] ——, *RTCGA.clinical: Clinical datasets from The Cancer Genome Atlas Project*, 2022, R package version 20151101.28.0.

[10] H. Iuchi, T. Matsutani, K. Yamada, N. Iwano, S. Sumi, S. Hosoda, S. Zhao, T. Fukunaga, and M. Hamada, "Representation learning applications in biological sequence analysis," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 3198–3208, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037021002208