Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

GIFT: Guided and Interpretable Factorization for Tensors - Applications to Human Cancer Analytics

Jungwoo Lee [1†], Sejoon Oh [1†], and Lee Sael [2*]

[1]Computer Science and Engineering, Seoul National University, Seoul, Korea,
[2]Computer Science, The State University of New York (SUNY), Inceon, Korea, and
† These authors contributed equally to this work.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Given cancer genome data with auxiliary gene set information, how can we extract significant relations between cancers, gene sets, and genes? How can we devise a tensor factorization method which produces interpretable factor matrices while maintaining the decomposition quality and speed?
Method: We propose **GIFT**, a **G**uided and **I**nterpretable **F**actorization for **T**ensors. GIFT provides interpretable factor matrices by encoding prior knowledge as a regularization term in its objective function.
Results: Experiment results demonstrate that GIFT produces interpretable factorizations with high scalability and accuracy, while other methods lack interpretability. We apply GIFT to the PANCAN12 dataset, and GIFT reveals significant relations between cancers, gene sets, and genes, such as influential gene sets for specific cancer (e.g., interferon-gamma response gene set for ovarian cancer) or relations between cancers and genes (e.g., BRCA cancer ↔ APOA1 gene and OV, UCEC cancers ↔ BST2 gene).
Availability: The code and datasets used in the paper are available at https://github.com/leesael/GIFT.
Contact: sael@cs.stonybrook.edu
Supplementary information: Supplementary material is available at https://github.com/leesael/GIFT.

1 Introduction

Given cancer genome data and gene set information, how can we discover gene sets or genes closely related to cancer in an integrative and interpretable way? Understanding relations between cancer and genes is the main problem in a genetic analysis of cancer since a gene mutation is a direct cause of cancer, thus the relations give us a clue for a treatment of cancer. While a vast amount of research on a genetic analysis of cancer has been conducted [2, 21, 22, 27], the Cancer Genome Atlas (TCGA) research network published a PanCan12 dataset in 2013 [36], which includes genomic information of 12 tumor types. The dataset has boosted many genetic cancer analytics [3, 13, 26].

Many real-world data are modeled as multi-dimensional arrays, which are called tensors. For example, movie rating and network traffic data are represented as 3-order tensors with (movie - user - time) and (source IP - destination IP - time) triples, respectively. As the PanCan12 dataset contains information about 14,591 genes of 4,993 patients from 5 different experiments, the dataset is expressed as a 3-order tensor in a form of (patient - gene - experiment; experiment value).

Table 1. Comparison of our proposed method GIFT and the other algorithms. GIFT produces interpretable results while maintaining high accuracy and scalability. However, P-Tucker and Silenced-TF are lack of interpretability or accuracy, respectively.

Method	P-Tucker [25]	Silenced-TF	**GIFT**
Interpretability		✓	✓
Accuracy	✓		✓
Scalability	✓	✓	✓

Modeling data into a tensor is widely used since it enables us to find meaningful patterns from the data by using a tool called tensor factorization (TF), which decomposes a given tensor into factor matrices and a core tensor. Applications of TF include anomaly detection from network traffic data [10], healthcare monitoring from sensor data [35], and fraud detection from social network data. Moreover, there have been various approaches which use TF on biomedical data [7, 16].

Most real-world tensors including the PanCan12 dataset are partially observed rather than fully observed; the former tensors contain many missing values, while the latter ones assume all entries are filled with values. Therefore, we concentrate on TF methods [6, 25, 28] which assume input tensors are partially observed.

In addition to the PanCan12 dataset, we have gene set data from MSigDB collections which are represented in a form of (gene - gene set). We apply a TF method on the PanCan12 dataset using the gene set data as prior knowledge. Many TF methods [6, 8, 15] exploit prior knowledge or guiding information to obtain high-quality factorizations or intended latent patterns. However, factor matrices produced by existing methods are hard to interpret due to their density and unclear value distributions.

Developing an interpretable TF method is essential for analyzing its resultant factors more effectively; poor interpretability makes it hard to discover latent patterns. The main challenge is to make factor matrices interpretable while preserving the TF quality such as training error. Our goal is to devise an interpretable TF method for partially observed tensors exploiting prior knowledge, while preserving the quality and scalability.

In this paper, we propose GIFT (Guided and Interpretable Factorization for Tensor), a TF method which provides interpretable factor matrices for partially observed tensors. GIFT utilizes prior knowledge as a regularization term during its factorizations. The guided regularization makes a clear distinction between values of factor matrices, which enhances interpretability of GIFT. We apply GIFT on the PanCan12 dataset with gene set data and discover notable relations between genes, gene sets, and cancer. Table 1 shows a comparison of GIFT, Silenced-TF, and P-Tucker [25], where Silenced-TF is a naive version of GIFT and P-Tucker is an existing TF method for partially observed tensors. GIFT provides interpretable factor matrices while keeping high accuracy and scalability, while the other two methods cannot meet both aspects.

Our main contributions are as follows.

- **Method.** We propose GIFT (Guided and Interpretable Factorization for Tensors). GIFT offers interpretable factor matrices by encoding prior knowledge.
- **Experiments** GIFT produces interpretable factorizations with high scalability and accuracy while the other methods are lack of accuracy or interpretability.
- **Discovery.** We apply GIFT to human cancer analytics using the PanCan12 dataset. We successfully discover significant relations between genes, gene sets, and cancer, as summarized in Table 4.

The rest of the paper is organized as follows. In Section 2, we give preliminaries on tensors and tensor factorizations. We explain our proposed method GIFT and other algorithms in Section 3. After presenting our experiment results in Section 4, we describe our discoveries on the PanCan12 dataset using GIFT in Section 5. We conclude in Section 6.

2 Preliminaries

In this section, we describe preliminaries of a tensor and its factorization methods. Table 2 summarizes symbols used in this paper.

2.1 Tensor

A tensor is a multi-dimensional array which is a generalization of a matrix and a vector. A mode or a way indicates each axis of a tensor, and an order is the number of modes or ways. We denote a tensor using boldface Euler script letters (e.g., $\mathcal{X}$). A tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is an N-order tensor which has N modes whose lengths are from I_1 to I_N. A vector and a matrix are regarded as a 1- and a 2-order tensor, respectively. We denote a matrix and a vector using boldface uppercase (e.g., $\mathbf{X}$) and lowercase letters (e.g., $\mathbf{x}$), respectively. The i_1th row of $\mathbf{A}$ is denoted by $\mathbf{a}_{i_1:}$, and the i_2th column of $\mathbf{A}$ is denoted by $\mathbf{a}_{:i_2}$.

Table 2. Table of symbols.

Symbol	Definition
$\mathcal{X}$	tensor (Euler script, bold letter)
$\mathbf{X}$	matrix (uppercase, bold letter)
x	scalar (lower case, italic letter)
N	order (number of modes) of a tensor
I_n, J_n	dimensionality of the nth mode of input and core tensor
$\mathbf{A}^{(n)}$	nth factor matrix ($\in \mathbb{R}^{I_n \times J_n}$)
$a^{(n)}_{i_n j_n}$	(i_n, j_n)th entry of $\mathbf{A}^{(n)}$
Ω	set of observable entries of $\mathcal{X}$
$\lvert\Omega\rvert, \lvert\mathcal{G}\rvert$	number of observable entries of input and core tensor
λ	regularization parameter for factor matrices
$\lvert\lvert \bullet \rvert\rvert_F$	Frobenius norm
$*$	element-wise multiplication
$\circ$	outer product
$\times_n$	n-mode product

2.2 Tensor Factorization

Among many tensor decomposition methods, we use Tucker factorization [32, 9] methods, which allows us to discover not only latent concepts but also relations between the concepts hidden in tensors [25]. Other factorization methods such as CP are described in [17]. Tucker factorization decomposes a given tensor $\mathcal{X}$ into a core tensor $\mathcal{G}$ and factor matrices $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}$, as defined in Definition 1.

Definition 1. ***(Tucker factorization)*** *Given a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, the Tucker factorization of rank $(J_1, \cdots, J_N)$ finds a core tensor $\mathcal{G} \in \mathbb{R}^{J_1 \times \cdots \times J_N}$ and factor matrices $\mathbf{A}^{(1)} \in \mathbb{R}^{I_1 \times J_1}, \cdots, \mathbf{A}^{(N)} \in \mathbb{R}^{I_N \times J_N}$, which minimize the following objective function* (1).

$$\begin{aligned}
&\mathcal{L}(\mathcal{G}, \mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}) \\
&= ||\mathcal{X} - \sum_{\forall (j_1, \ldots, j_N) \in \mathcal{G}} \mathcal{G}_{j_1, \ldots, j_N} (\mathbf{a}^{(1)}_{j_1} \circ \cdots \circ \mathbf{a}^{(N)}_{j_N})||^2_F \\
&= \sum_{\forall (i_1, \ldots, i_N) \in \mathcal{X}} \left(\mathcal{X}_{(i_1, \ldots, i_N)} - \sum_{\forall (j_1, \ldots, j_N) \in \mathcal{G}} \mathcal{G}_{(j_1, \ldots, j_N)} \prod_{n=1}^{N} a^{(n)}_{i_n j_n} \right)^2
\end{aligned} \tag{1}$$

Note that Equation (1) assumes missing entries of $\mathcal{X}$ as zeros. Each column vector of a factor matrix generally represents each different concept. A higher value in a vector indicates that the corresponding element is highly related to the concept. Assuming a given tensor is movie rating data with (movie - user - time) triples, then a column vector in a movie-factor matrix can have a concept such as a horror or comic genre.

2.3 Partially Observable Tensor Factorization

Many real-world tensors have missing values in it (i.e. partially observed). Applying standard Tucker factorization methods to the data triggers highly inaccurate results since they regard missing entries as zeros. Partially observable tensor factorization methods focus only on observed entries to tackle this problem, and a partially observable Tucker factorization is defined as follows.

Definition 2. ***(Partially Observable Tucker Factorization)*** *Given a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with observable entries Ω, a partially observable Tucker factorization of rank $(J_1, \cdots, J_N)$ finds a core tensor $\mathcal{G} \in \mathbb{R}^{J_1 \times \cdots \times J_N}$ and factor matrices $\mathbf{A}^{(1)} \in \mathbb{R}^{I_1 \times J_1}, \cdots, \mathbf{A}^{(N)} \in$*

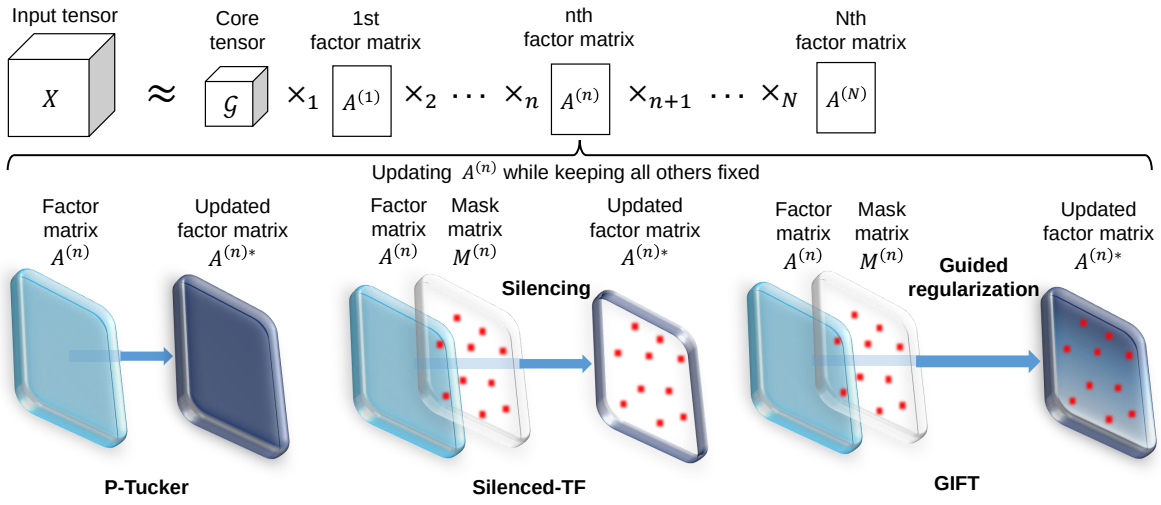


Fig. 1. An overview of GIFT and the other methods. All methods are based on a Tucker factorization, which decomposes a given tensor into a core tensor and factor matrices. Tucker decomposition methods generally update a factor matrix at a time, while fixing all the other factor matrices. P-Tucker updates a factor matrix using a normal L_2 regularization, and the results of P-Tucker are hard to interpret. On the other hand, Silenced-TF and GIFT employ prior information encoded as mask matrices for higher interpretability. Silenced-TF silences some values of a factor matrix as zeros according to the mask matrix. Meanwhile, GIFT updates those entries through a guided regularization instead of fixing them as zeros.

$\mathbb{R}^{I_N \times J_N}$ *which minimize the following objective function.*

$$\mathcal{L}(\mathcal{G}, \mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}) = \sum_{\forall (i_1,\ldots,i_N) \in \mathcal{X}} \left(\mathcal{X}_{(i_1,\ldots,i_N)} - \sum_{\forall (j_1,\ldots,j_N) \in \mathcal{G}} \mathcal{G}_{(j_1,\ldots,j_N)} \prod_{n=1}^{N} a^{(n)}_{i_n j_n} \right)^2 + \lambda \left(\sum_{n=1}^{N} ||\mathbf{A}^{(n)}||_F^2 \right) \tag{2}$$

Note that λ denotes a regularization parameter for factor matrices, and we used L_2-regularization to prevent overfitting, which has been widely used in recommender systems [18, 29].

3 Proposed Method

In this section, we describe our proposed method GIFT. We introduce a baseline approach P-Tucker, naive interpretable approach Silenced-TF and then our proposed GIFT.

3.1 Overview

Given prior knowledge, how can we enhance interpretability of factor matrices? The main challenge is to devise a method which employs prior knowledge and produces interpretable factors while maintaining the decomposition quality and speed. We define a factor matrix is interpretable if it is easy to distinguish intended elements from unintended ones in the same factor matrix. We encode prior knowledge into mask matrices $\mathbf{M}$ which are binary matrices with 0 or 1 and have the same dimensionality to factor matrices $\mathbf{A}$. The proposed naive Silenced-TF silences all the unintended elements into 0. Silenced-TF can cause large reconstruction error due to many silenced elements. We propose our GIFT which selectively impose regularization on unintended elements and not on intended ones thus making distinction easier while guaranteeing the reconstruction error.

3.2 Baseline Approach: P-Tucker

Among many Tucker factorization methods [30, 24, 11], P-Tucker [25] shows the best scalability and accuracy for partially observable tensors by focusing on observed entries of the tensors. The objective function of P-Tucker is the same to Equation (2), and P-Tucker uses a row-wise alternating least squares (ALS) to minimize the loss function. In detail, P-Tucker first chooses a row of a factor matrix to be updated while fixing all the others, and it computes three intermediate data $\delta^{(n)}_{(i_1,\ldots,i_N)}$, $\mathbf{B}^{(n)}_{i_n}$, and $\mathbf{c}^{(n)}_{i_n:}$ defined as follows. Notice that $\mathbf{I}_{J_n}$ is a $J_n \times J_n$ identity matrix. $\delta^{(n)}_{(i_1,\ldots,i_N)}$ is a length J_n vector whose jth entry is

$$\sum_{\forall (j_1 \ldots j_n = j \ldots j_N) \in \mathcal{G}} \mathcal{G}_{(j_1 \ldots j_n = j \ldots j_N)} \prod_{k \neq n} a^{(k)}_{i_k j_k}, \tag{3}$$

$\mathbf{B}^{(n)}_{i_n}$ is a $J_n \times J_n$ matrix whose (j_1, j_2)th entry is

$$\sum_{\forall (i_1,\ldots,i_N) \in \Omega^{(n)}_{i_n}} \delta^{(n)}_{(i_1,\ldots,i_N)}(j_1) \delta^{(n)}_{(i_1,\ldots,i_N)}(j_2), \tag{4}$$

and $\mathbf{c}^{(n)}_{i_n:}$ is a length J_n vector whose jth entry is

$$\sum_{\forall (i_1,\ldots,i_N) \in \Omega^{(n)}_{i_n}} \mathcal{X}_{(i_1,\ldots,i_N)} \delta^{(n)}_{(i_1,\ldots,i_N)}(j). \tag{5}$$

Using the above intermediate data, P-Tucker updates a row $a^{(n)}_{i_n:}$ by an update rule $\mathbf{c}^{(n)}_{i_n:} \times [\mathbf{B}^{(n)}_{i_n} + \lambda \mathbf{I}_{J_n}]^{-1}$. After updating factor matrices, P-Tucker calculates reconstruction error by the following rule.

$$\sqrt{\sum_{\forall (i_1,\ldots,i_N) \in \Omega} \left(\mathcal{X}_{(i_1,\ldots,i_N)} - \sum_{\forall (j_1,\ldots,j_N) \in \mathcal{G}} \mathcal{G}_{(j_1,\ldots,j_N)} \prod_{n=1}^{N} a^{(n)}_{i_n j_n} \right)^2} \tag{6}$$

If the error converges or the maximum iteration is reached, P-Tucker stops iterations and performs QR decompositions to orthogonalize factor matrices and update a core tensor accordingly. Note that [25] suggests full details and proofs of the update process.

3.3 Interpretable Approaches: Silenced-TF and GIFT

Silenced-TF. Although P-Tucker presents high scalability and accuracy while decomposing given tensors, it is hard to interpret the results of P-Tucker since there are no distinctions between values of intended and unintended entries. Hence, we propose Silenced-TF which provides interpretable factors and bridges between P-Tucker and GIFT. Silenced-TF

Algorithm 1 GIFT

Input: A tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with observable entries Ω, mask matrices $\mathbf{M}^{(1)}, \cdots, \mathbf{M}^{(N)}$, rank $(J_1, \cdots, J_N)$, and a regularization parameter λ.
Output: A core tensor $\mathcal{G}$ and factor matrices $\mathbf{A}^{(1)}, \cdots \mathbf{A}^{(N)}$.
1: initialize $\mathcal{G}$ and $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}$ randomly
2: **repeat**
3: **for** $n = 1, \cdots, N$ **do**
4: **for** $i_n = 1, \cdots, I_n$ **do**
5: calculate intermediate data δ, $\mathbf{B}^{(n)}_{i_n}$, and $\mathbf{c}^{(n)}_{i_n:}$ by (3) – (5)
6: calculate $\mathbf{D}^{(n)}_{i_n}$, where its (j_n, j_n)th entry is $\mathbf{M}^{(n)}_{i_n, j_n}$
7: update a row $a^{(n)}_{i_n:}$ by $\mathbf{c}^{(n)}_{i_n:} \times [\mathbf{B}^{(n)}_{i_n} + \lambda \mathbf{D}^{(n)}_{i_n}]^{-1}$
8: **end for**
9: **end for**
10: compute reconstruction error by (6)
11: **until** convergence criterion is met
12: **for** $n = 1 \ldots N$ **do**
13: $\mathbf{A}^{(n)} \rightarrow \mathbf{Q}^{(n)} \mathbf{R}^{(n)}$
14: $\mathbf{A}^{(n)} \leftarrow \mathbf{Q}^{(n)}$
15: $\mathcal{G} \leftarrow \mathcal{G} \times_n \mathbf{R}^{(n)}$
16: **end for**

literally silences uninteresting or unnecessary parts of factor matrices and updates the rest of parts using the same algorithm of P-Tucker. To be more specific, given the mask matrices, Silenced-TF only updates an entry $a^{(n)}_{i_n j_n}$ when the corresponding masking element $m^{(n)}_{i_n j_n}$ is 0, and Silenced-TF makes an entry $a^{(n)}_{i_n j_n}$ as a zero when $m^{(n)}_{i_n j_n}$ is 1.

Objective function of Silenced-TF. The revised loss function of Silenced-TF is given as follows.

$$\begin{aligned} &\underset{\mathcal{G}, \mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}}{\text{minimize}} && \mathcal{L}(\mathcal{G}, \mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}, \mathbf{M}^{(1)}, \cdots, \mathbf{M}^{(N)}) \\ &\text{subject to} && a^{(n)}_{i_n j_n} = 0 \; when \; m^{(n)}_{i_n j_n} = 1 \end{aligned} \tag{7}$$

Implementation of Silenced-TF. The only difference between Silenced-TF and P-Tucker is updating a row of a factor matrix. Silenced-TF updates $a^{(n)}_{i_n j_n}$ when $m^{(n)}_{i_n j_n} = 0$; otherwise, it replaces the entry with a zero, while P-Tucker updates all entries in the row regardless of the mask matrix.

GIFT. The main problem of Silenced-TF is its low factorization accuracy due to many zeros in its factor matrices. Therefore, it is crucial to develop a method which offers both interpretable factorizations and high accuracy. Our proposed method GIFT tackles the problem by employing selective regularizations of factor matrices. GIFT gives penalties with a strength λ to unintended entries of factor matrices during the update process. Thus, GIFT makes a distinction between the values of intended and unintended ones. Moreover, the accuracy of GIFT is similar to P-Tucker since GIFT does not fix the values of unintended ones as zeros.

Objective function of GIFT. GIFT encodes mask matrices $\mathbf{M}^{(n)}$ into its objective function as a regularization term. The specific loss function of GIFT is given by the following Equation (8).

$$\begin{aligned} &\mathcal{L}(\mathcal{G}, \mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(N)}, \mathbf{M}^{(1)}, \cdots, \mathbf{M}^{(N)}) = \\ &\sum_{\forall (i_1, \ldots, i_N) \in \Omega} \left(\mathcal{X}_{(i_1, \cdots, i_N)} - \sum_{\forall (j_1, \ldots, j_N) \in \mathcal{G}} \mathcal{G}_{(j_1, \cdots, j_N)} \prod_{n=1}^{N} a^{(n)}_{i_n j_n} \right)^2 \\ &+ \lambda \left(\sum_{n=1}^{N} ||\mathbf{M}^{(n)} * \mathbf{A}^{(n)}||^2 \right) \end{aligned} \tag{8}$$

The main difference between P-Tucker and GIFT is an existence of mask matrices $\mathbf{M}^{(n)}$ in a regularization term. GIFT uses $\mathbf{M}^{(n)} * \mathbf{A}^{(n)}$ instead of just $\mathbf{A}^{(n)}$, where $*$ denotes an element-wise multiplication. Through the specially-designed regularization, GIFT fully exploits prior knowledge encoded in $\mathbf{M}^{(n)}$. Compared to Silenced-TF, GIFT shows flexibility regarding the updates of unintended entries. Instead of fixing them as zeros, GIFT imposes regularizations on them, which tend to make the values smaller, but not normally zeros.

Implementation of GIFT. Algorithm 1 describes how GIFT updates given factor matrices. When GIFT updates a row $a^{(n)}_{i_n:}$ (line 6), it requires a diagonal matrix D which reflects masking information (line 5), while P-Tucker uses an identity matrix $\mathbf{I}_{J_n}$. The other parts of GIFT are the same to that of P-Tucker.

Table 3. Summary of datasets used for experiments. M: million, K: thousand.

Dataset	Order	Dimensionality	Observable Entries
PANCAN12 tensor	3	(4993, 14591, 5)	180M
Sampled-PANCAN12	3	(4993, 14591, 5)	36−144M
Mask matrix $\mathbf{M}^{(2)}$	2	(14591, 50)	7K

4 Experiment

In this section, we describe experimental results of GIFT compared to Silenced-TF and P-Tucker. We aim to answer the following questions.

[Q1] Interpretability: How interpretable are factor matrices produced by GIFT and the other methods? (Section 4.3)

[Q2] Accuracy: How accurately do GIFT and the other methods factorize a given tensor and predict missing entries of the tensor? (Section 4.4)

[Q3] Scalability: How well do GIFT and the other methods scale up with respect to the number of observed entries of a tensor? (Section 4.5)

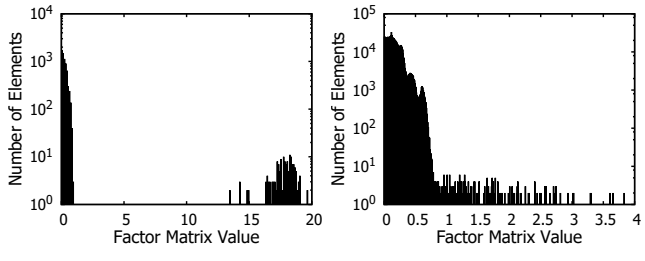
We first describe our datasets and experimental settings. After that, we answer the above questions in Sections 4.3 to 4.5,

4.1 Datasets

We use PANCAN12 and Hallmark gene set data from MSigDB collections as an input tensor and a mask matrix, respectively. Table 3 summarizes the data we used in this paper. PANCAN12 data is represented as a 3-order tensor in a form of (patient - gene - experiment type; experiment value), and we generate a mask matrix $\mathbf{M}^{(2)}$ in a form of (gene - gene set) using the Hallmark data, which contain 50 important gene sets. In $\mathbf{M}^{(2)}$, each column corresponds to each gene set, and we give $\mathbf{M}^{(2)}_{i_n j_n}$ 0 (intended) if a gene i_n is contained in a gene set j_n; otherwise, we set them to 1 (unintended). For other mask matrices, we fill them with zeros. Note that patients in PANCAN12 data are classified by 12 disjoint cancer groups.

4.2 Experimental Settings

GIFT and other methods are implemented in C with OpenMP and Eigen libraries. We run our experiments on a single machine with 20 cores / 40 threads, equipped with an Intel Xeon E5-2630 v4 2.2GHz CPU and 512GB RAM. We set the default rank as $(30 \times 50 \times 2)$. In reporting running times,



(a) Value distribution of intended entries derived by GIFT

(b) Value distribution of unintended entries derived by GIFT

Fig. 2. Distributions of values in a gene-factor matrix derived by GIFT (λ = 10). In a set of intended entries, there is a huge gap between large and small values. There are also few large values in a set of unintended entries. Hence, GIFT makes a distinction in each group.

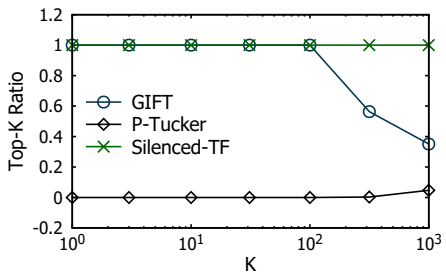


Fig. 3. Top-K ratios for GIFT and other methods. A high top-K ratio implies that the corresponding method is interpretable. P-Tucker shows the worst top-K ratio since it cannot distinguish intended and unintended entries. the ratios of GIFT decrease when K become larger since GIFT includes important unintended entries in the top-K set.

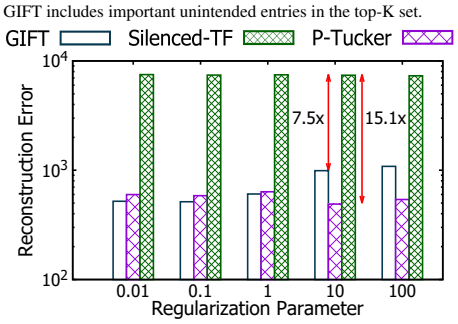


Fig. 4. Reconstruction error of GIFT and other methods. Silenced-TF is highly inaccurate due to many zeros in its factor matrix, while GIFT and P-Tucker present a high accuracy.

we use the average elapsed time per iteration, not the total running time. Notice that we use absolute values of factor matrices for all experiments.

4.3 Interpretability

We regard a method as interpretable if it makes a distinction between the values of intended and unintended entries. In other words, an interpretable method ought to make values of intended entries high and vice versa. An entry is intended when it belongs to one of the gene sets and vice versa. As presented in Figure 2, GIFT makes a huge gap between values of some intended entries and the others. Silenced-TF also makes a distinction by fixing the values of unintended entries as zeros. However, P-Tucker cannot distinguish the values of intended and unintended entries. (refer to the supplementary material for the results of Silenced-TF and P-Tucker). Additionally, we use a top-K ratio as a metric of interpretability. A top-K ratio indicates how many intended entries are included in total top-K entries (in descending order by values) of a gene-factor matrix, which is defined as follows.

$$\text{Top-K ratio R } (0 \leq R \leq 1) = \frac{\text{number of intended entries in top-K}}{K} \tag{9}$$

A high top-K ratio implies that a method increases values of intended entries, or decreases values of unintended ones, or the both. Thus, a method with a high top-K ratio meets a criterion for high interpretability. Figure 3 illustrates top-K ratios of GIFT and other methods regarding K. P-Tucker shows the worst top-K ratios for all K since it treats intended and unintended entries in the same way. Although Silenced-TF exhibits the highest top-K ratios for all K by silencing the unintended entries, Silenced-TF cannot extract important unintended entries which are closely related to intended entries since their values are all set to zeros. Meanwhile, the top-K ratio of GIFT is the highest until $K \leq 10^2$ and decreases rapidly when $K \geq 10^2$ since the GIFT includes important unintended entries in the top-K set. Overall, Silenced-TF and GIFT provide interpretable factorizations with respect to distributions of values in a factor matrix and top-K ratios.

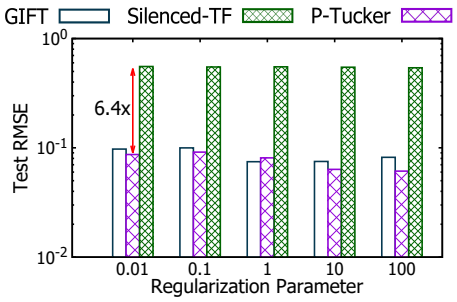


Fig. 5. Test RMSE of GIFT and other methods. Silenced-TF is highly inaccurate due to many zeros in its factor matrix, while GIFT and P-Tucker present a high accuracy.

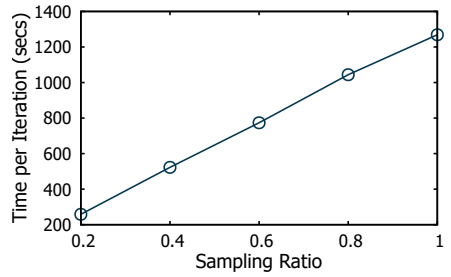


Fig. 6. Scalability of GIFT with respect to the number of observable entries in the tensor. As the number of observed entries increases, a running time of GIFT increases proportionally.

4.4 Accuracy

We use two evaluation metrics—reconstruction error and test root mean square error (RMSE)—to measure the accuracy of GIFT and other methods. Specifically, reconstruction error indicates an accuracy of a factorization as given in Equation (6), and test RMSE implies how accurately a method predicts missing entries of a tensor. Notice that we split the PANCAN12 tensor into training/test data with a ratio of 9 to 1 to measure the test RMSE. As illustrated in Figures 4 and 5, Silenced-TF exhibits the worst accuracy due to many zeros in a silenced factor matrix. The reconstruction error and test RMSE of Silenced-TF are $15.1\times$ and $6.4\times$ higher than that of P-Tucker when $\lambda = 10$ and $\lambda = 0.01$, respectively. While P-Tucker shows the best accuracy in most cases, GIFT presents relatively small accuracy loss compared to that of Silenced-TF; in particular, test RMSE of GIFT is slightly higher or even better than that of P-Tucker.

4.5 Scalability

We vary the number of observable entries by randomly sampling 20%, 40%, 60%, 80%, and 100% from the PANCAN12 tensor. As shown in Figure 6, GIFT scales near linearly in terms of the number of observable entries. We omit results of P-Tucker and Silenced-TF since they present similar scalability to that of GIFT (refer to the supplementary material).

5 Discovery

In this section, we describe our discoveries on (cancer - gene sets), (gene sets - genes), and (cancer - genes) relations hidden in the PANCAN12 dataset by interpreting results of GIFT. We set λ to 10 for the discoveries.

5.1 Cancer - Gene sets

Given specific cancer, which gene set is the most relevant to the cancer? We first explain our discovery procedure and introduce several examples of (cancer - gene sets) relations found by GIFT.

We first compute an influence of each gene set on a patient and extract top-k important gene sets for the patient. After that, we aggregate all top-k gene sets of patients suffering from the given cancer and derive top-k relevant gene sets to the cancer by choosing top-k frequent gene

Table 4. Discoveries on the PANCAN12 dataset found by our proposed method GIFT. GIFT extracts significant gene sets (e.g., TGF beta signaling and interferon-gamma response) and notable relations between cancer, gene sets, and genes (e.g., BRCA cancer ↔ APOA1 gene and OV, UCEC cancers ↔ BST2 gene). Many biological research results substantiate the retrieved relations, which are described in a gene description column. (*: important gene, $^-$: unimportant gene, $^+$: not included in a gene set, but related)

Cancer	Gene set	Genes	Gene value	Gene Description
HNSC, LUAD, LUSC, BLCA	TGF beta signaling	SKIL*	0.5146	The SKIL gene encodes the SNON, negative regulators of TGF-beta signaling [31].
		FKBP1A*	0.4692	The FKBP1A gene interacts with a type I TGF-beta receptor.
		LEFTY2*	0.2925	The LEFTY2 gene encodes a secreted ligand of the TGF-beta family of proteins.
GBM	Angiogenesis	PF4*	0.5049	The PF4 gene inhibits cell proliferation and angiogenesis in vitro and in vivo [5].
		VCAN*	0.4500	The VCAN gene encodes a protein involving in cell adhesion, and angiogenesis [37].
		LPL$^-$	0.0429	The LPL gene encodes lipoprotein lipase [34].
BRCA	Estrogen response late	IL17RB*	0.3807	The IL17RB gene is important in development and progression of breast cancer [2].
		TFF3*	0.3640	The TFF3 gene promotes invasion and migration of breast cancer [22].
		PTGER3$^-$	0.0200	The encoded protein by the PTGER3 gene is related to digestion and nervous system.
	Bile acid metabolism	APOA1*	0.3973	The LPL gene encodes lipoprotein lipase, an enzyme which hydrolyzes lipoprotein. Higher APOA1 level is known to be related to increased breast cancer risk [21].
OV, UCEC	Interferon-gamma response	IRF7*	0.5727	The IRF7 gene encodes interferon regulatory factor 7.
		BST2*	0.4986	High levels of BST2 have been identified in ovarian cancer [27].
		SSPN$^-$	0.0983	The SSPN gene is associated with a skeletal muscle membrane [20].
	Apoptosis	CASP8AP2$^+$	1.4708	The CASP8AP2 gene is associated with apoptosis of leukemic lymphoblasts [12]. The protein encoded by the CASP8AP2 plays a regulatory role in Fas-mediated apoptosis [14].
READ, COAD	Protein secretion	STX7*	0.4854	The STX7 gene controls vesicle trafficking events involved in cytokine secretion [1].
KIRC, LAML	Mitotic spindle	LATS1*	0.4913	The LATS1 gene binds phosphorylated zyxin and moves it to the mitotic spindle.

sets in aggregations. In detail, $\mathbf{a}^{(1)}_{i:}$ is a latent feature of i-th patient, and $\mathbf{G} = (\sum_{i=1}^{I_3} \mathcal{G}_{::i})/I_3$ is a relation between gene sets and columns of a patient-factor matrix. Then, we use $\tilde{\mathbf{a}}^{(1)}_{i:} = \mathbf{a}^{(1)}_{i:}\mathbf{G}$ as an influence of each gene set on the i-th patient. The j-th element of $\tilde{\mathbf{a}}^{(1)}_{i:}$ indicates the influence of j-th gene set on the i-th patient. We extract top-k most important gene sets for each patient by selecting top-k highest values in $\tilde{\mathbf{a}}^{(1)}_{i:}$. Finally, we count the frequency of gene sets appeared in the top-k gene sets of all patients having the given cancer. We regard the most frequent gene set as the most relevant one to the given cancer.

The first and second columns of Table 4 show (cancer - gene sets) relations discovered by GIFT. For breast cancer (BRCA), GIFT considers 'Estrogen response late' and 'Bile acid metabolism' gene sets closely related to breast cancer. It is well known that the estrogen plays a key role in the occurrence of breast cancer [4] while the relation to 'Bile acid metabolism' gene set seems unnatural. However, Murray et al. [23] reveal that patients with breast cancer have significantly low fecal bile acid concentration than that of controlled patients. For ovarian cancer (OV), a relation to the 'Interferon-gamma response' gene set is supported by Wall. et al [33]. They show that interferon-gamma causes apoptosis in human epithelial ovarian cancer. The 'TGF beta signaling' gene set is frequent among many types of cancer including Head and Neck Squamous Cell Carcinoma (HNSC), Lung adenocarcinoma (LUAD), Lung Squamous, Cell Carcinoma (LUSC), and Bladder carcinoma (BLCA). The reason is that the Transforming growth factor-β (TGF-β) gene set is a tumor suppressor which affects many types of human cancers [19].

5.2 Gene sets - Genes

Given a gene set, which genes are important or unimportant within it? Is there a gene not included but related to the gene set? A high value in the gene-factor matrix indicates that the corresponding gene is highly related to the corresponding gene set. We sort the genes in each column of the gene-factor matrix in descending order by their value and inspect high-value genes for each gene set.

We show how the discovered (gene sets - genes) relations are supported by biological facts with examples. The second and third columns of Table 4 show (gene sets - genes) relations retrieved by GIFT. The SKIL gene in the 'TGF beta signaling' gene set is identified to be important by GIFT. The SKIL gene encodes a protein which antagonizes TGF-β signaling [31]. The PF4 gene in the 'Angiogenesis' gene set, reported to be important by GIFT, is known as an inhibitor of cell proliferation and angiogenesis [5]. The IRF7 gene in the 'Interferon-gamma response' gene set is also identified to be important, and the gene encodes interferon regulatory factor 7. The LPL gene in the 'Angiogenesis' gene set is unimportant according to the discovery result of GIFT. The LPL gene encodes lipoprotein lipase, an enzyme which hydrolyzes lipoprotein [34], thus it has low relatedness to angiogenesis. GIFT also reports the CASP8AP2 gene is closely related to the 'Apoptosis' gene set although the gene is not included in the gene set. The CASP8AP2 gene is associated with apoptosis of leukemic lymphoblasts in reality [12].

5.3 Cancer - Genes

Given specific cancer, which genes affect the cancer most? We suggest (cancer - genes) relations by combining two relations (cancer - gene sets) and (gene sets - genes) discovered by GIFT.

The first and third columns of Table 4 show (cancer - genes) relations found by GIFT. We regard gene sets in the second column of the table as bridges for (cancer - genes) relations. We deduce the IL17RB and TFF3 genes are significant to breast cancer since the genes are both important for the 'Estrogen response late' gene set and the gene set is the most relevant one to breast cancer. In reality, IL17RB is crucial in development and progression of breast cancer in effect [2]. Moreover, May et al. reveal that the TFF3 gene promotes invasion and migration of breast cancer [22]. GIFT also finds that the APOA1 gene in the 'Bile acid metabolism' gene set is highly related to breast cancer. High levels of APOA1 are known to be related to increased breast cancer risk [21]. In the case of ovarian cancer, GIFT asserts a strong relation to the BST2 gene. High levels of BST2 have been identified in ovarian cancer [27].

6 Discussion and Conclusion

In this paper, we propose GIFT, a guided and interpretable factorization method for tensors. GIFT provides interpretable factor matrices by

encoding prior knowledge through selective regularizations. Experiment results demonstrate that GIFT produces interpretable factorizations with high scalability and accuracy, while other methods are lack of accuracy or interpretability. In practice, we apply GIFT to human cancer analytics using the PANCAN12 dataset and successfully identify important relations between cancers, gene sets, and genes. For instance, GIFT suggests influential gene sets for specific cancer (e.g., interferon-gamma response gene set for ovarian cancer). In addition, GIFT provides remarkable relations between cancers and genes (e.g., BRCA cancer $\leftrightarrow$ APOA1 gene and OV, UCEC cancers $\leftrightarrow$ BST2 gene). Furthermore, GIFT is able to extract out-of-the-box relations, which are not given in prior information. Specifically, in Hallmark gene set data, a CASP8AP2 gene was not included in a gene set about apoptosis. However, GIFT elicits a relation between the gene and gene set, which is an acknowledged relation by the papers [12, 14]. Although GIFT is a general framework for interpretable tensor factorizations with prior knowledge, its accuracy or discoveries might not be better than that of other techniques for fine-grained tasks (e.g., focusing on single cancer) or datasets with no prior knowledge. Even in those cases, GIFT is still effective since the quality of a decomposition does not depend on the granularity of tasks and GIFT is transformed into P-Tucker if there is no prior information. Future works of GIFT include utilizing various types of regularizations such as Lasso, applications to higher-order data and finding complex relations between multi-axes, or implementing GIFT in distributed platforms for a large amount of human genome data.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2017R1A1A1A05001325). The ICT at Seoul National University provided research facilities for this study.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2017R1A1A1A05001325). The ICT at Seoul National University provided research facilities for this study.

References

[1]A. Achuthan, P. Masendycz, J. A. Lopez, T. Nguyen, D. E. James, M. J. Sweet, J. A. Hamilton, and G. M. Scholz. Regulation of the endosomal snare protein syntaxin 7 by colony-stimulating factor 1 in macrophages. *Molecular and cellular biology*, 28(20):6149–6159, 2008.

[2]V. Alinejad, S. Dolati, M. Motallebnezhad, and M. Yousefi. The role of il17b-il17rb signaling pathway in breast cancer. *Biomedicine & Pharmacotherapy*, 88:795–803, 2017.

[3]J. Anaya, B. Reon, W.-M. Chen, S. Bekiranov, and A. Dutta. A pan-cancer analysis of prognostic genes. *PeerJ*, 3:e1499, 2016.

[4]A. Ao, B. J. Morrison, H. Wang, J. A. López, B. A. Reynolds, and J. Lu. Response of estrogen receptor-positive breast cancer tumorspheres to antiestrogen treatments. *PLoS One*, 6(4):e18810, 2011.

[5]A. Bikfalvi. Platelet factor 4: an inhibitor of angiogenesis. In *Seminars in thrombosis and hemostasis*, volume 30, pages 379–385. Copyright© 2004 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA., 2004.

[6]D. Choi, J.-G. Jang, and U. Kang. Fast, accurate, and scalable method for sparse coupled matrix-tensor factorization. *arXiv preprint arXiv:1708.08640*, 2017.

[7]A. Cichocki. Tensor decompositions: a new concept in brain data analysis? *arXiv preprint arXiv:1305.0395*, 2013.

[8]I. Davidson, S. Gilpin, O. Carmichael, and P. Walker. Network discovery via constrained tensor analysis of fmri data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 194–202. ACM, 2013.

[9]L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[10]T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, editors. *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*. ACM, 2006.

[11]M. Filipović and A. Jukić. Tucker factorization with missing data with application to low-n-rank tensor completion. *Multidimensional systems and signal processing*, 26(3):677–692, 2015.

[12]C. Flotho, E. Coustan-Smith, D. Pei, S. Iwamoto, G. Song, C. Cheng, C.-H. Pui, J. R. Downing, and D. Campana. Genes contributing to minimal residual disease in childhood acute lymphoblastic leukemia: prognostic significance of casp8ap2. *Blood*, 108(3):1050–1057, 2006.

[13]K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[14]Y. Imai, T. Kimura, A. Murakami, N. Yajima, K. Sakamaki, and S. Yonehara. The CED-4-homologous protein FLASH is involved in Fas-mediated activation of caspase-8 during apoptosis. *Nature*, 398(6730):777–785, Apr 1999.

[15]B. Jeon, I. Jeon, L. Sael, and U. Kang. Scout: Scalable coupled matrix-tensor factorization-algorithm and discoveries. In *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*, pages 811–822. IEEE, 2016.

[16]S. Kim, L. Sael, and H. Yu. A mutation profile for top-k patient search exploiting gene-ontology and orthogonal non-negative matrix factorization. *Bioinformatics*, 31(22):3653–3659, 2015.

[17]T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009.

[18]Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[19]M. Kretzschmar. Transforming growth factor-β and breast cancer: transforming growth factor-β/smad signaling defects and cancer. *Breast Cancer Research*, 2(2):107, 2000.

[20]K. A. Lapidos, R. Kakkar, and E. M. McNally. The dystrophin glycoprotein complex. *Circulation research*, 94(8):1023–1031, 2004.

[21]L. J. Martin, O. Melnichouk, E. Huszti, P. W. Connelly, C. V. Greenberg, S. Minkin, and N. F. Boyd. Serum lipids, lipoproteins, and risk of breast cancer: a nested case-control study using multiple time points. *JNCI: Journal of the National Cancer Institute*, 107(5), 2015.

[22]F. E. May and B. R. Westley. Tff3 is a valuable predictive biomarker of endocrine response in metastatic breast cancer. *Endocrine-related cancer*, 22(3):465–479, 2015.

[23]W. Murray, A. Blackwood, K. Calman, and C. MacKay. Faecal bile acids and clostridia in patients with breast cancer. *British journal of cancer*, 42(6):856–860, 1980.

[24]J. Oh, K. Shin, E. E. Papalexakis, C. Faloutsos, and H. Yu. S-hot: Scalable high-order tucker decomposition. In *WSDM*, 2017.

[25]S. Oh, N. Park, L. Sael, and U. Kang. Scalable Tucker Factorization for Sparse Tensors - Algorithms and Discoveries. *arxiv*, 2017.

[26]N. Riaz, P. Blecua, R. S. Lim, R. Shen, D. S. Higginson, N. Weinhold, L. Norton, B. Weigelt, S. N. Powell, and J. S. Reis-Filho. Pan-cancer analysis of bi-allelic alterations in homologous recombination dna repair genes. *Nature Communications*, 8(1):857, 2017.

[27]Y. Shigematsu, N. Oue, Y. Nishioka, N. Sakamoto, K. Sentani, Y. Sekino, S. Mukai, J. Teishima, A. Matsubara, and W. Yasui. Overexpression of the transmembrane protein bst-2 induces akt and erk phosphorylation in bladder cancer. *Oncology Letters*, 14(1):999–1004, 2017.

[28]K. Shin, L. Sael, and U. Kang. Fully scalable methods for distributed tensor factorization. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):100–113, 2017.

[29]K. Shin, L. Sael, and U. Kang. Fully scalable methods for distributed tensor factorization. *TKDE*, 29(1):100–113, 2017.

[30]S. Smith and G. Karypis. Accelerating the tucker decomposition with compressed sparse tensors. In *Europar*, 2017.

[31]A. C. Tecalco-Cruz, M. Sosa-Garrocho, G. Vázquez-Victorio, L. Ortiz-García, E. Domínguez-Hüttinger, and M. Macías-Silva. Transforming growth factor-β/smad target gene skil is negatively regulated by the transcriptional cofactor complex snon-smad4. *Journal of Biological Chemistry*, 287(32):26764–26776, 2012.

[32]L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[33]L. Wall, F. Burke, C. Barton, J. Smyth, and F. Balkwill. Ifn-Î³ induces apoptosis in ovarian cancer cells in vivo and in vitro. 9:2487–96, 08 2003.

[34]H. Wang and R. H. Eckel. Lipoprotein lipase: from gene to obesity. *American Journal of Physiology-Endocrinology and Metabolism*, 297(2):E271–E288, 2009.

[35]X. Wang, C. Yang, and S. Mao. Tensorbeat: Tensor decomposition for monitoring multi-person breathing beats with commodity wifi. *CoRR*, abs/1702.02046, 2017.

[36]J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[37]T. N. Wight. Versican: a versatile extracellular matrix proteoglycan in cell biology. *Current opinion in cell biology*, 14(5):617–623, 2002.