Vol.2021-BIO-65 No.2
2021/3/11

カーネルテンソル分解を用いた 教師なし学習による変数選択法 ～ バイオインフォマティクスへの応用 ～

田口 善弘[1,a)]

概要：変数の数に比べてサンプル数が少ないいわゆる *large p small n* 問題については，多くの研究が為されている．だがいわゆるゲノム科学の分野ではこの比が極端であり，遺伝子の数（＝変数の数＝ p) が数万個であるのに対して被験者の数（＝サンプル数＝ n）が数個の場合さえあり $p/n \sim 10^3$ であることも稀ではない．このような極端な場合にはいわゆる *large p small n* 問題に対して提案された多くの方法が無力である場合が多い．我々はこの問題に対処するため「主成分分析あるいはテンソル分解を用いた教師無し学習による変数選択法」を提案し，この１０年程の間に多くのバイオインフォマティクスの分野の研究に応用してきた．しかし，この方法は純粋に線形代数の範囲内の方法であり，教師無し学習であるためにチューニングパラメーターもなく，手法がうまく行かない場合には解析そのものを諦める以外になかった．今回，我々はこの問題を解決するために同手法のカーネル化に成功した．これにより同法は非線形関係を考慮できるようになり大きくその適応範囲が拡大したということができるのでその内容について報告する．

キーワード：テンソル分解，カーネルトリック，変数選択法，教師無し学習

Kernel tensor decomposition based unsupervised feature extraction – Applications to bioinformatics –

Abstract: A lot of research has been done on the so-called *large p small n* problem, where the number of samples is small compared to the number of variables. In the so-called field of genome science, however, this ratio is extreme, with the number of genes (= number of variables = p) being tens of thousands while the number of subjects (= number of samples = n) is even a few, and $p/n \sim 10^3$ is not uncommon. In such extreme cases, many of the methods proposed for the so-called *large p small n*problem are often ineffective. We have proposed "Principal component analysis and tensor decomposition based unsupervised feature extraction" to deal with this problem. In the past decade, we have applied this method to many researches in the field of bioinformatics. However, this method is purely within the scope of linear algebra, and since it is unsupervised learning, there is no tuning parameter, and if the method does not work, there is no choice but to give up on the analysis itself. In order to solve this problem, we have developed a kernel version of the method. In this paper, we report on how we succeeded in kernelizing the method to solve this problem, which enables the method to take nonlinear relationships into account and greatly expands its application range.

Keywords: Tensor decomposition, Kernel trick, Feature selection, Unsupervised learning

1. はじめに

いわゆる *large p small n* 問題，あるいは，高次元統計の問題は現在でも解析が難しい問題として立ちはだかっている．なんらかの予測問題として考えた場合にはサンプル数

[1] 中央大学理工学部物理学科
Department of Physics, Chuo Univeristy, Tokyo 112-8551, Japana
[a)] tag@granular.com

（＝ n ＝条件の数）が変数の数（＝ p）より少なくなってしまうため，どんな予測でも１００％の精度で当たってしまういわゆる過学習を避けることができない．この様な問題を避けるためには次元を下げるために新たな合成変数を作る，あるいは疎性モデリングなどの方法を使って少数個の変数を選択するなどの方法が提案されてきた．しかし，これらの方法は *large p small n* 問題とは言っても $p/n \sim 10$ のオーダーが精々であり，そもそも p/n が極端に大きな場合はあまり想定されていない．現実問題として p/n 比が極端に大きい場合は稀であり，例外的な問題として扱われていた嫌いがある．実際，最近の機械学習でも，多数回の試行による大規模な学習が可能であることは暗黙の了解であり，p/n 比が大きい場合は最初から考察の対象から除かれており，このような問題を扱う場合には転移学習などの枠組みで類似したデータセットでの大規模な事前学習をすることで乗り切る方向性が主流であった．

これに対し，いわゆるゲノム科学の分野では最初から p/n 比が大きいことが前提である．ヒトゲノムの塩基配列は３０億程度であるが，人類の人口はそもそも７０億程度である．全人類の相当程度をサンプリングしない限り，$p/n \sim 10$ の様なことは想定出来ない．そもそも人類は大型生物としては極端に個体数が多い生命種であり，多くの人間程度，あるいは人間より大きい生物種の総個体数は数千である場合も多い．例えばトラの個体数は２０世紀初頭には１０万頭程度であったと言われるが，現在は数千頭までその数を減じている．しかし，トラのゲノムサイズは人間とさして変わらず，仮に１０万頭いたとしても p/n が極端に大きいという現実は変わらない．従って，ゲノム科学を扱うにはどうしても p/n 比が大きい場合の効果的な解決方法が（ゲノム科学に特化したものであっても）必要である．

この問題に正面から向き合って来たかどうかはなはだ怪しい．実際，遺伝子（変数）ごとに独立性を仮定して t 検定，あるいはその派生形であるな統計的な検定を行って，遺伝子数を考慮した多重比較補正を行う方法が一般的であるが，この方法はサンプル数によってその統計の検定能力が大きく左右されるため，サンプル数が増えてしまうと検定を通過する変数の数が多くなりすぎて，p/n を減ずるという当初の目的がそもそも実行できなくなったり，逆に十分なサンプル数がないと有意差がある変数が一個もないという問題に直面する．ゲノム科学の場合は，この問題を他の基準，例えば Fold Change（二群の差を比較するときの比の絶対値）がある値（例えば２）以下の変数は最初から解析から除く，などの方法で対処してきた．プラクティカルにはこの対処法で問題がないため放置されているが決して根本的な解決が為されたとは言えない．

我々はこの問題を解決するために従来から「主成分分析およびテンソル分解を用いた教師なし学習による変数選択法」 [1] を提案してこの問題の解決にあたってきた．この方法は非常に広範なバイオインフォマティクスの問題に有効であることが示されたが，線形の教師無し学習による方法であるため，ある特定の問題でうまく行かない場合には，チューニングパラメーターなどがそもそも存在しないため，手法の適用そのものを諦めるしかないのが常であった．この様な限界を突破するため，我々は同手法のカーネル化（＝カーネルトリックを用いることができるようにする一般化）を試みた．その結果，非常に簡便な方法で，同手法は容易にカーネル化できることが判明した [2]．本稿ではその結果と応用例について説明する．

2. 数学的な定式化

「カーネルテンソル分解を用いた教師なし学習による変数選択法」（以下，提案手法）について説明する前に従来の「主成分分析を用いた教師なし学習による変数選択法」及び「テンソル分解を用いた教師なし学習による変数選択法」について簡単に説明する．これらの目的は極端な *large p small n* 問題において効果的な変数選択を行うことである．

2.1 主成分分析を用いた教師なし学習による変数選択法

$x_{ij} \in \mathbb{R}^{N\times M}$ は i 番目の遺伝子の j 番目のサンプルの遺伝子発現量であるとする．ここで $\sum_i x_{ij} = 0, \sum_i x_{ij} = N$ となる様に規格化されていると仮定する．ここでグラム行列 $x_{ii'} = \sum_j x_{ij}x_{i'j} \in \mathbb{R}^{N\times N}$ を対角化

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

することで遺伝子 i の第 ℓ 主成分得点 $u_{\ell i} \in \mathbb{R}^{N\times N}$ 及び固有値 λ_ℓ を計算する．サンプル j の第 ℓ 主成分負荷量 $v_{\ell j} \in \mathbb{R}^{M\times M}$ は

$$v_{\ell j} = \sum_i x_{ij}u_{\ell i} \tag{2}$$

で計算できる．余談であるが $v_{\ell j}$ は $x_{jj'} = \sum_i x_{ij}x_{ij'} \in \mathbb{R}^{M\times M}$ の固有ベクトルでもある．なぜなら

$$\sum_{j'} x_{jj'}v_{\ell j'} = \sum_{j'}\sum_i x_{ij}x_{ij'}\sum_{i'} x_{i'j'}u_{\ell i'} \tag{3}$$

$$= \sum_{ii'} x_{ij}\sum_{j'} x_{ij'}x_{i'j'}u_{\ell i'} \tag{4}$$

$$= \sum_i x_{ij}\sum_{i'} x_{ii'}u_{\ell i'} = \sum_i x_{ij}\lambda_\ell u_{\ell i} = \lambda_\ell v_{\ell j} \tag{5}$$

「主成分分析を用いた教師なし学習による変数選択法」ではまず $v_{\ell j}$ の j 依存性に注目し，どの ℓ が生物学的に重要かを判断する．例えば，j が患者と健常者からなっており，患者と健常者で発現差がある遺伝子を探したい，と言った場合には $v_{\ell j}$ のうち患者と健常者の二群で差がある主成分負荷量をまず探し，これに対応する主成分得点 $u_{\ell i}$ が多重ガウス分布していることを仮定して（帰無仮説）累積 χ 二乗分布 $P_{\chi^2}[> x]$ を用いて

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

という式で P 値を付与し（σ_ℓ は標準偏差），この P 値を BH 基準 [1] で多重比較補正してある閾値（経験的には 0.01 を閾値とするとよい結果が得られるようである）として閾値以下の補正 P 値を付与された i を選択する．

2.2 テンソル分解を用いた教師なし学習による変数選択法

前節の方法はそのままテンソル分解に直接拡張できる．$x_{ijk} \in \mathbb{R}^{N\times M\times K}$ は i 番目の遺伝子の j 番目のサンプルの k 番目の臓器の遺伝子発現量であるとする．ここで $\sum_i x_{ijk} = 0, \sum_i x_{ijk}^2 = N$ となる様に規格化されていると仮定する．Higher order singular value decomposition (HOSVD) [1] を用いて

$$x_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{M}\sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \tag{7}$$

を得る．$G \in \mathbb{R}^{N\times M\times K}$ はコアテンソルで積 $u_{\ell_1 i}u_{\ell_2 j}u_{\ell_3 k}$ の寄与の重みを表現している．$u_{\ell_1 i} \in \mathbb{R}^{N\times N}, u_{\ell_2 j} \in \mathbb{R}^{M\times M}, u_{\ell_3 k} \in \mathbb{R}^{K\times K}$ は特異値ベクトルで x_{ijk} の i,j,k 依存性を表現している．「主成分分析を用いた教師なし学習による変数選択法」の時と同じように，まず $u_{\ell_2 j}$ と $u_{\ell_3 k}$ の j,k 依存性に注目し，どの ℓ_2, ℓ_3 が生物学的に重要であるかを判断する．$u_{\ell_2 j}$ については「主成分分析を用いた教師なし学習による変数選択法」と同じように考えて選び，$u_{\ell_3 k}$ については例えば特定の臓器 k に特異的に値が大きい特異値ベクトルを探すことにする．「主成分分析を用いた教師なし学習による変数選択法」の時と違い，サンプル依存性を記述する $u_{\ell_2 j}, u_{\ell_3 k}$ と遺伝子 i（変数）選択に用いる $u_{\ell_1 i}$ との間に 1 対 1 対応がないので選択された ℓ_2, ℓ_3 を固定したうえで大きな $|G(\ell_1\ell_2\ell_3)|$ を与えるような（つまり x_{ijk} への寄与が大きい）ℓ_1 を選択することで i の選択に用いる特異値ベクトル $u_{\ell_1 i}$ を選ぶ．この $u_{\ell_1 i}$ を用いて「主成分分析を用いた教師なし学習による変数選択法」の時と同じように

$$P_i = P_{\chi^2}\left[> \sum_{\ell_1} \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}} \right)^2 \right] \tag{8}$$

という式で P 値を付与し，この P 値を BH 基準 [1] で多重比較補正してある閾値（経験的には 0.01 を閾値とするとよい結果が得られるようである）として閾値以下の補正 P 値を付与された i を選択する．

「テンソル分解を用いた教師なし学習による変数選択法」は単に「主成分分析を用いた教師なし学習による変数選択法」の主成分分析をテンソル分解に置き換えただけではなく「主成分分析を用いた教師なし学習による変数選択法」ではできなかったことが可能になる．それは異なったデータの統合解析，バイオインフォマティクスの分野で言えばマルチオミックス解析と呼ばれる解析が可能になる．$x_{ij} \in \mathbb{R}^{N\times M}$ は i 番目の遺伝子の j 番目のサンプルの遺伝子発現量，$x_{kj} \in \mathbb{R}^{K\times M}$ は k 番目のゲノム部位の j 番目のサンプルのプロモーターメチル化（という遺伝子発現量とは別の観測量），であるとしよう．注意すべきは i と k には対応関係がないので，同じサンプル j に見た目上は全く独立な（ただし生物学的にはなんらかの関係があるであろう）j と k という二種類の観測を行ったことになっていることである．ここでメチル化が高いと発現量が減ることが知られているので

- 遺伝子発現プロファイルが患者と健常者で差がある．
- プロモーターメチル化が患者と健常者で差がある．
- 遺伝子発現プロファイルとプロモーターメチル化の間に（負の）相関があるペアを探したい．

という 3 条件を満たす (i,k) のペアを知りたいということになる．この 3 つの要請は独立なので基本的には 3 種類の独立な検定をおこなうべきだが，問題は大きい．3 つの検定を独立に行った結果，どれくらいオーバーラップがあるか不安である．特に三番目の要請は j と k のペアについての要請なので $(N\times K)/2$ 回の非常に多数回の検定を行わないといけないため，多重比較補正を考えるとかなり小さな P 値が得られないと有意な相関ではないとみなされてしまうことは考えられるが $M \ll N, K$ であるために，M は小さく，大きな相関が仮に得られても十分な P 値を得るのは難しい．

ここで

$$x_{ijk} = x_{ij}x_{kj} \in \mathbb{R}^{N\times M\times K} \tag{9}$$

という風に行列からテンソルを構成してこのテンソル分解を (7) 式と同じように計算することを考える．この場合，サンプル依存性を表現しているのは $u_{\ell_2 j}$ だけなので，まず，「主成分分析を用いた教師なし学習による変数選択法」の時と同じように生物学的に重要な ℓ_2 を選択することになる．そして選んだ ℓ_2 を固定したうえで大きい $|G(\ell_1\ell_2\ell_3)|$ を持っている ℓ_1, ℓ_3 を選択することになる．そして（8）式と同じようにして i に P 値を，

$$P_k = P_{\chi^2}\left[> \sum_{\ell_3} \left(\frac{u_{\ell_3 k}}{\sigma_{\ell_3}} \right)^2 \right] \tag{10}$$

で k に P 値を付与する．多重比較補正と i,k の選択基準はいままでと同じである．

この方法には 1 つ欠点がある．もとのデータ量は $(N+K)\times M$ であるがこれが $N\times M\times K$ に増大してしまう点である．この点を改良するために以下の様な近似的な計算も提案した．

$$x_{ik} = \sum_j x_{ijk} \in \mathbb{R}^{N\times K} \tag{11}$$

この x_{jk} をＨＯＳＶＤすれば（実際には行列なのでただの

ＳＶＤになってしまうが）

$$x_{ik} = \sum_{\ell} \lambda_{\ell} u_{\ell i} u_{\ell k} \tag{12}$$

という形で i,k に付与される特異値ベクトルを（近似的に）計算できる．これでは j に付与された特異値ベクトルが計算できないがそれは近似的に

$$u^{(i)}_{\ell j} = \sum_{i} x_{ij} u_{\ell i} \tag{13}$$

$$u^{(k)}_{\ell j} = \sum_{k} x_{kj} u_{\ell k} \tag{14}$$

という式で計算が可能である．これでは j に付与される特異値ベクトルが２種類求まってしまうがそれは x_{ij}, x_{kj} の j 依存性をそれぞれ表現していると思えばよいであろう．$u^{(i)}_{\ell j}, u^{(k)}_{\ell j}$ の j 依存性を見ることで生物学的に意味がある ℓ を同定した後，対応する $u_{\ell i}, u_{\ell k}$ を用いて i,k にＰ値を付与して多重比較補正し，i,k を選ぶ手順はこれまでと同じようにすればよい．

2.3 カーネルテンソル分解を用いた教師無し学習による変数選択法

ここまでの方法にはいくつか欠点がある．まず，チューニングパラメータがないので，うまく行かなかった場合に改善の余地がない．また，「テンソル分解を用いた教師なし学習による変数選択法」による統合解析（マルチオミックス解析）の場合には，データ量が $(N+K)\times M$ から $N\times M\times K$ あるいは $N\times K$ に増えてしまう．$M \ll N, K$ を想定しているのでこれはかなり致命的である．この２つの問題はテンソル分解をカーネル化することでかなりの程度まで改善する．

まずテンソル分解をカーネル化するには内積の導入が必要である．カーネルトリックは内積をカーネルで置き換えてそのまま線形演算をする手法だからだ．このため

$$x_{jkj'k'} = \sum_{i} x_{ijk} x_{ij'k'} \in \mathbb{R}^{M\times K\times M\times K} \tag{15}$$

という計算（以下これを線形カーネルと呼ぶ）を導入すると

$$\begin{aligned}
x_{jkj'k'} &= \sum_{i} \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \\
&\times \sum_{\ell'_1=1}^{N} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} G(\ell'_1\ell'_2\ell'_3) u_{\ell'_1 i} u_{\ell'_2 j'} u_{\ell'_3 k'} \qquad (16)\\
&= \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} \sum_{\ell'_1=1}^{N} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} G(\ell_1\ell_2\ell_3) G(\ell'_1\ell'_2\ell'_3) \\
&\times \left(\sum_{i} u_{\ell_1 i} u_{\ell'_1 i} \right) u_{\ell_2 j} u_{\ell_3 k} u_{\ell'_2 j'} u_{\ell'_3 k'} \qquad (17)\\
&= \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} \sum_{\ell'_1=1}^{N} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} G(\ell_1\ell_2\ell_3) G(\ell'_1\ell'_2\ell'_3) \\
&\times \delta_{\ell_1\ell'_1} u_{\ell_2 j} u_{\ell_3 k} u_{\ell'_2 j'} u_{\ell'_3 k'} \qquad (18)\\
&= \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} \left(\sum_{\ell_1=1}^{N} G(\ell_1\ell_2\ell_3) G(\ell_1\ell'_2\ell'_3) \right) \\
&\times u_{\ell_2 j} u_{\ell_3 k} u_{\ell'_2 j'} u_{\ell'_3 k'} \qquad (19)\\
&= \sum_{\ell_2=1}^{M} \sum_{\ell_3=1}^{K} \sum_{\ell'_2=1}^{M} \sum_{\ell'_3=1}^{K} G'(\ell_2\ell_3\ell'_2\ell'_3) \\
&\times u_{\ell_2 j} u_{\ell_3 k} u_{\ell'_2 j'} u_{\ell'_3 k'} \qquad (20)
\end{aligned}$$

（但し，$G'(\ell_2\ell_3\ell'_2\ell'_3) = \sum_{\ell_1=1}^{N} G(\ell_1\ell_2\ell_3) G(\ell_1\ell'_2\ell'_3)$）を得るので，$x_{jkj'k'}$ にＨＯＳＶＤを適用することで $u_{\ell_2 j}, u_{\ell_3 k}$ は計算できることが期待される．（15）式は内積の形をしているのでこれを線形カーネルと見なすことで任意の正定値カーネルに置き換えればカーネルトリックを用いたテンソル分解を容易に実行できる．正定値カーネルとしては

$$x_{jkj'k'} = \exp\left\{ -\alpha \sum_{i} \left(x_{ijk} - x_{ij'k'} \right)^2 \right\} \tag{21}$$

$$= \left(1 + \sum_{i} x_{ijk} x_{ij'k'} \right)^{d} \tag{22}$$

などが有名である．（21）式は Radial base function (RBF) カーネル，（22）式は多項式カーネルと呼ばれている．

テンソル分解のカーネル化は統合解析（マルチオミックス解析）でデータ量を減らすことに役立つ．

$$x^{(i)}_{jj'} = \sum_{i} x_{ij} x_{ij'} \in \mathbb{R}^{M\times M} \tag{23}$$

$$x^{(k)}_{jj'} = \sum_{k} x_{kj} x_{kj'} \in \mathbb{R}^{M\times M} \tag{24}$$

を導入しておき

$$x_{jj'} = \sum_{j''} x^{(i)}_{jj''} x^{(k)}_{j''j'} \in \mathbb{R}^{M\times M} \tag{25}$$

を計算してからこれにＨＯＳＶＤを適用すれば，

$$x_{jj'} = \sum_{\ell_2=1}^{M} \sum_{\ell'_2=1}^{M} G(\ell_2\ell'_2) u^{(i)}_{\ell_2 j} u^{(k)}_{\ell'_2 j'} \tag{26}$$

を得ることができ，ここから

$$u_{\ell_2 i} = \sum_{j} x_{ij} u^{(i)}_{\ell_2 j} \tag{27}$$

$$u_{\ell_2 k} = \sum_{j} x_{kj} u^{(k)}_{\ell_2 j} \tag{28}$$

を得られる．あとの流れは同じなので省略する．大切なことは N,K といった大きな数の次元の行列計算は不要になり $\mathbb{R}^{M\times M}$ の行列の演算しかなくなったことだ．これでデータ量がふえてしまうことを避けることができるようになった．

但し，以上の計算は線形カーネルの場合しか使えない．RBF カーネルや多項式カーネルの場合は上記の式で $u^{(i)}_{\ell_2 j}$ や $u^{(k)}_{\ell_2 j}$ から $u_{\ell_2 i}$ や $u_{\ell_2 k}$ を計算することはできない．$u_{\ell_2 i}$

や $u_{\ell_2 k}$ を計算できなければこれらを用いて i や k に P 値を付与することもできない．P値が計算できなければ i や k を選択することも叶わない．その場合は，非常に計算量は増えてしまうが，以下の様にするしか方法がない．特定の i や k を順番に除いてからカーネルを計算し，ＨＯＳＶＤを適用して $u^{(i)}_{\ell_2 j}$ や $u^{(k)}_{\ell_2 j}$ を計算する．そして $u^{(i)}_{\ell_2 j}$ や $u^{(k)}_{\ell_2 j}$ の「劣化度」が高い順に i や k を選ぶ．ここで劣化度とは，$u^{(i)}_{\ell_2 j}$ や $u^{(k)}_{\ell_2 j}$ の j 依存性の問題である．j が患者と健常者からなっているのであれば，$u^{(i)}_{\ell_2 j}$ や $u^{(k)}_{\ell_2 j}$ に t 検定などの統計検定を適用し，有意さがより大きく減ずるものから順に選択していく．これはより j 依存性に意味がある i や k を除けば計算された $u^{(i)}_{\ell_2 j}$ や $u^{(k)}_{\ell_2 j}$ の j 依存性もより大きく減ずるだろうという予想の元に考えられた方法である．この方法は非常に原始的であるし，計算量的にも $x_{jj'}$ のＨＯＳＶＤを K 回，あるいは，N 回繰り返す必要があり，データ量は増えないものの計算時間が某大になってしまうという欠点もあるが，下記に見るようにかなりうまく行くことが経験的には分かっている．

3. 結果

以下では「主成分分析を用いた教師なし学習による変数選択法」や「テンソル分解を用いた教師なし学習による変数選択法」との比較という形で提案手法の有効性を検証していく．

3.1 スイスロール

カーネルトリック導入のもっとも大きな動機は線形ではとらえきれない非線形性の扱いであった．提案手法がデータの非線形性をとらえることができるかを確認しよう．ここではスイスロールと呼ばれる，データの非線形性のテストにもっともよくつかわれるデータを採用する．$x_{ijk} \in \mathbb{R}^{N\times 3\times 10}$ を

$$p_i = -1 + \frac{2i}{N}, 1 \le i \le N \tag{29}$$

$$x_{i1k} = p_i \cos(2\pi p_i) \tag{30}$$

$$x_{i2k} = -1 + \varepsilon_{ijk} \tag{31}$$

$$x_{i3k} = p_i \sin(2\pi p_i) \tag{32}$$

のように定義する．これは１０通り（$1 \le k \le 10$）の異なった乱数から作成されたスイスロールのバンドルに相当する．以下では $N = 10^3$ とし，また，$\sum_i x_{ijk} = 0, \sum_i x^2_{ijk} = N$ という規格化を行ってから計算を進めるものとする．x_{ij1} を図 1(A) に示す．x_{ij1} に SVD を適用しただけでは，図 1(B) の様に全く非線形をとられられていない．x_{ijk} に HOSVD を適用すると向きは揃うもの非線形性は全く捉えれていない（図 1(C)）．しかし，RBF カーネルに HOSVD を適用した図 1(D) ではわずかながら非線形を反映した埋め込みになっていることが分かる．

3.2 *large p small n*

スイスロールの場合には提案手法はうまく機能することが分かった．しかし，これらは３次元空間に１０００個の点を配置した例であり *large p small n* とは真逆の $n \gg p$ の状況である．提案手法のターゲットはそもそも *large p small n* 問題への適用であるから，この場合にうまく行かなくては意味がない．そこで

$$x_{ijk} \sim \begin{cases} \mathcal{N}(\mu,\sigma) & i \le N_1 \le N, j,k \le \frac{M}{2} \\ \mathcal{N}(0,\sigma) & \text{otherwise} \end{cases} \tag{33}$$

の様なテンソル $x_{ijk} \in \mathbb{R}^{N\times M\times M}$ を用意しよう．$N \gg M$ のように設定すれば *large p small n* 問題となる．このテンソルは $i \le N_1$ の i において二群になっているが，j,k がそれぞれ二群に分かれているのに j,k で張られる二次元の空間でも二群になっているという意味でちょっと意地悪な設定になっている．j,k の二変数があり，それらがそれぞれ二群に分かれていれば 2×2 で４群に分かれていると想定するのが普通だからだ．

以下では RBF カーネル，多項式カーネル，線形カーネルを用いた提案手法の性能比較を行った．また，比較のために，RBF カーネルと多項式カーネルについてはカーネル主成分分析も行った．カーネル主成分分析を行う場合には x_{ijk} はアンフォルディングして $x_{i(jk)} \in \mathbb{R}^{N\times M^2}$ の様な N 行 M^2 列の行列として扱った．それぞれのパフォーマンス評価は得られた特異値ベクトルが二群に分かれている程度のよさで評価した．提案手法の場合には $u_{\ell j}u_{\ell k}$ に $j,k \le \frac{M}{2}$ の $\frac{M^2}{4}$ 個の (j,k) の組とそれ以外の $\frac{3}{4}M^2$ 個の (j,k) の組の２群に対して t 検定を行い，M 個の P 値を得る．次にこれらを多重比較補正し，もっとも小さな P 値を記録する．このプロセスを乱数を変えて１００回行い，Ｐ値の幾何平均を提案手法の評価値とする（勿論，Ｐ値の幾何平均が小さい方がよいとする）．カーネル主成分分析の場合は，N 行 M^2 列の行列を扱うのでＰ値が M^2 個求まるという以外は同じである．表 1 はその結果である．まず，気づくことは多重比較補正しない場合にはカーネル主成分分析と提案手法には差が無いということである．多重比較補正して初めて提案手法の方がよくなる．これはカーネル主成分分析ではＰ値が M^2 個計算されるのに対して，提案手法ではＰ値が M 個しか計算されないためである．テンソル分解にすることで自由度が減ったことが有効に働いている．一方で提案手法どうしの比較では，非線形性を考慮できるはずのＲＢＦカーネルや多項式カーネルが線形カーネルを越えられていない．スイスロールの時は明確な差があったにも拘わらず，どうしてだろうか？これは以下の様に考えられる．スイスロールの時は点が１０００個もあったので図 1(A) に見られるような非線形の複雑な構造を表現できていた．しかし，今回は $M^2 = 36$ 個という必ずしも非線形性を十分に表現できない数の点であるため，非線

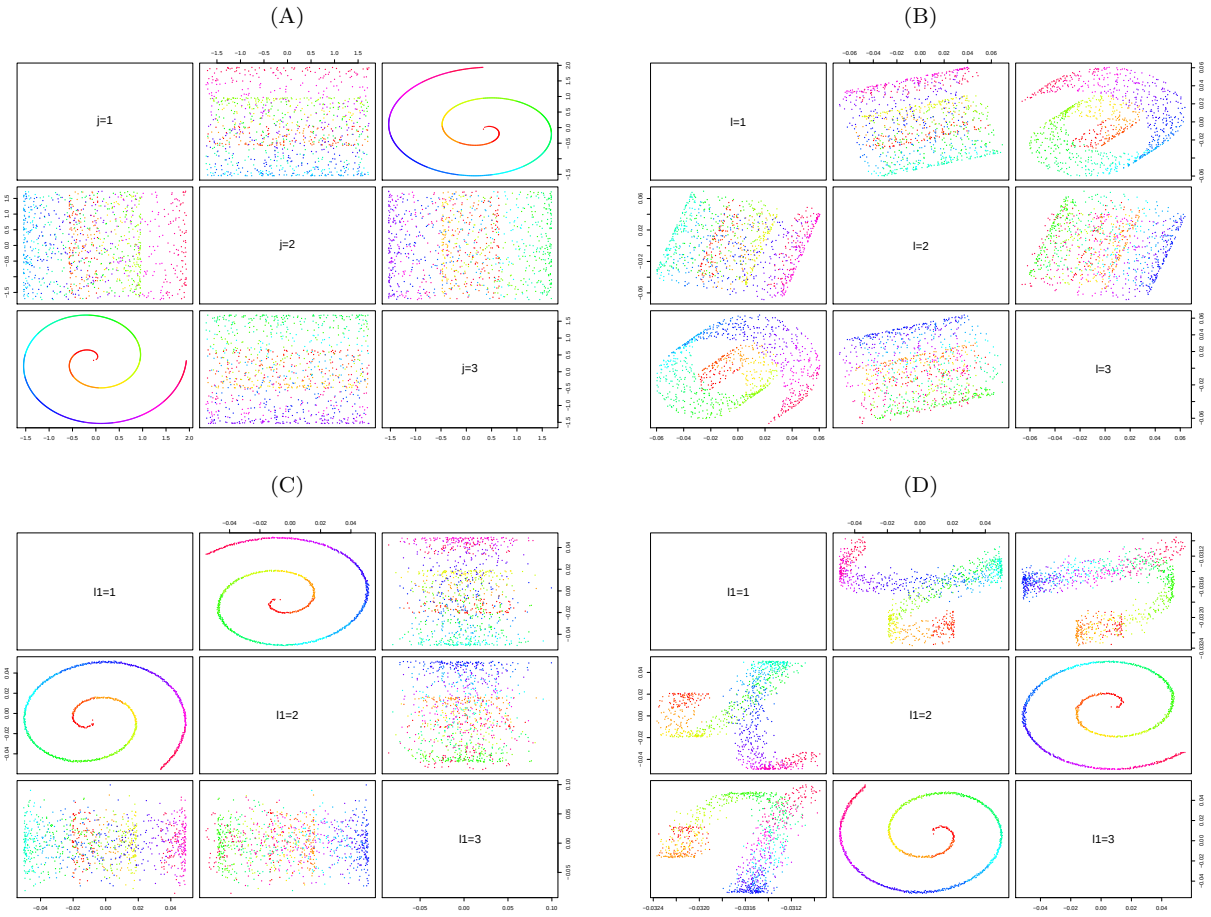

図 1 スイスロール．カラーグラデーションは i (1 to N) の向きを示す． (A) x_{ij1}, (B) $u_{\ell i}, 1 \leq \ell \leq 3$ SVD, (C) $u_{\ell_1 i}, 1 \leq \ell \leq 3$ HOSVD, (D) $u_{\ell_1 i}, 1 \leq \ell \leq 3$ RBF カーネル（$\alpha = 0.01$）に HOSVD を適用した場合．

Fig. 1 Swiss roll. The colors represent the direction of i (1 to N) in gradation. (A) x_{ij1}, (B) $u_{\ell i}, 1 \leq \ell \leq 3$ by SVD, (C) $u_{\ell_1 i}, 1 \leq \ell \leq 3$ by HOSVD, (D) $u_{\ell_1 i}, 1 \leq \ell \leq 3$ by kernel (RBF,$\alpha = 0.01$) based HOSVD.

表 1 提案手法，またはカーネル主成分分析を用いて得られた特異値ベクトルに t 検定を適用して得られた最小の P 値の幾何平均を計算した．P値は小さい方が性能が良い．$N = 1000, N_1 = 10, M = 6, \mu = 2, \sigma = 1$.

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

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

形カーネルが有効には働かなかったことが分かる．*large p small n* 問題における非線形性は微妙な問題であり，なかなか人工データでは拾いきれない．やはりどうしても現実のデータでの比較が必要だ．以下では現実の *large p small n* 問題に提案手法を適用してみよう．

3.3 COVID-19 のドラッグリポジショニング

COVID-19 は全世界で猛威を振るっている新型のコロナウィルスによる肺炎である．その感染力の強さと高齢者が集中的に死亡し，無症状でも感染力はあるという特異な性質のため，コロナの感染の広がりを抑えることが難しくなっている．一方で重症化する患者は感染者のごく少数（数パーセント以下）であるために，重症化を抑える薬さえ発見されれば，コロナの感染の広がりを抑えるために社会をシャットダウンして経済的なダメージを被るという問題をある程度解決できるために，治療薬の開発が急がれている．通常，創薬は１０年以上の時間がかかるプロセスであるがＣＯＶＩＤ－１９の場合は経済へのダメージが大きいため，より早期の創薬が望まれている．その場合，ゼロから薬を探索する通常の創薬では時間がかかりすぎるため，既知の薬の中から有望な薬を探すいわゆるドラッグリポジショニングが望まれている．

我々は先に，「テンソル分解を用いた教師なし学習による変数選択法」を用いて，ＣＯＶＩＤ－１９のインシリコドラッグリポジショニングを行っている [3]．手法の詳細については割愛するが，ＣＯＶＩＤ－１９の原因ウィルスである SARS-CoV-2 をヒトの肺臓由来の複数の培養細胞に感染させた際の遺伝子発現プロファイルに「テンソル分解を用いた教師なし学習による変数選択法」を適用し，感染細胞と比感染細胞の比較から，感染に際して大きく発現が変化する少数個の遺伝子を同定することが肝である．この研究では同定された遺伝子により大きな影響を与える化合物をデータベースから探索することで，ＣＯＶＩＤ－１９の治療に有効な化合物を推定した．推定された化合物の中には最近，ＣＯＶＩＤ－１９の予防や重症化阻止に有効であることが判明したイベルメクチンが含まれており，この方法の有効性が示された（２０２１年1月３０日ＴＢＳニュース「新型コロナ治療薬にイベルメクチン 南アフリカが限定的使用開始」２０２１年1月２６日日本経済新聞「イベルメクチン，都立病院で治験　コロナで東京都検討」）．

本節では提案手法を同じデータに用いることでよりよい結果が得られるかを検討する（詳細は原著論文 [2] 参照）．ここで何をよりよい結果とするかは難しいが，以下の様に考える．SARS-CoV-2 はウイルスであるため，単体では増殖できず，必ずヒトの細胞に侵入しなくてはならない．これを感染と呼ぶ．しかし，ウィルスのゲノムは小さく，感染に必要なタンパクを全て備えているわけではなく，ヒトのタンパクを利用する．先行研究で既にどのようなヒトのタンパクがウィルスのタンパクと相互作用するかは知られているので提案手法で選択した遺伝子が「テンソル分解を用いた教師なし学習による変数選択法」で選んだ遺伝子よりも，ウィルスと相互作用することが知られているヒトのタンパクとより大きく重なっていれば改善された，と見なすことにする．

解析対象のデータはテンソル $x_{ijkm} \in \mathbb{R}^{N\times 5\times 2\times 3}$ の形に整形されており，これは i 番目の遺伝子の j 番目の培養細胞の感染 $(k=1)$ ／非感染（$k=2$）における 3 つの Biological replicate の 1 つにおける発現量を表現している．これにＨＯＳＶＤを適用することで

$$x_{ijkm} = \sum_{\ell_1=1}^{5}\sum_{\ell_2=1}^{2}\sum_{\ell_3=1}^{m}\sum_{\ell_4=1}^{N} G(\ell_1\ell_2\ell_3\ell_4) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell_4 i} \quad (34)$$

を得た．$G(\ell_1\ell_2\ell_3\ell_4) \in \mathbb{R}^{5\times 2\times 3\times N}$ はコアテンソル，$u_{\ell_1 j} \in \mathbb{R}^{5\times 5}, u_{\ell_2 k} \in \mathbb{R}^{2\times 2}, u_{\ell_3 m} \in \mathbb{R}^{3\times 3}, u_{\ell_4 i} \in \mathbb{R}^{N\times N}$ は直交行列である特異値行列である．目的に叶った遺伝子を探すには培養細胞依存性がなく（つまり，$u_{\ell_1 j}$ は j によらない一定値をとる），感染非感染では差があり（つまり，$u_{\ell_2 1} = -u_{\ell_2 2}$ である），そして Biological Replicate の間で差がない（つまり，$u_{\ell_3 m}$ は m によらない一定値をとる）ような ℓ_1, ℓ_2, ℓ_3 の組を探す必要がある．結果的には $\ell_1 = \ell_3 = 1$ 及び $\ell_2 = 2$ がこの条件を満たすことが分かった．次に遺伝子選択に用いる $u_{\ell_4 i}$ を探す必要がある．これは $|G(1,2,1,\ell_4)|$ が最大になる ℓ_4 を選べばよいがこれは $\ell_4 = 5$ であった．そこでこれを用いて

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{5i}}{\sigma_5}\right)^2 \right] \quad (35)$$

の様な式で遺伝子 i にＰ値を付与する．これを BH 基準 [1] で多重比較補正して，0.01 以下の補正 P 値を付与された 163 遺伝子が選ばれた．これが先行研究 [3] の結果である．表 2 の「先行研究」の列に，163 遺伝子とウィルスと相互作用することが知られている遺伝子との一致度を示した．全てのウィルス遺伝子と相互作用するヒト遺伝子と非常に強く重なっている．

ここでは提案手法がこの先行研究の「非常に良い」結果を越えることができるかを考えよう．３種類の α $(1\times 10^{-2}, 1\times 10^{-4}, 1\times 10^{-6})$ のＲＢＦカーネルを使って

$$x_{jkmj'k'm'} = \exp\left\{ -\alpha \sum_i \left(x_{ijkm} - x_{ij'k'm'}\right)^2 \right\} \quad (36)$$

を計算し，これにＨＯＳＶＤを適用して

$$x_{jkmj'k'm'} = \sum_{\ell_1=1}^{5}\sum_{\ell_2=1}^{2}\sum_{\ell_3=1}^{3}\sum_{\ell'_1=1}^{5}\sum_{\ell'_2=1}^{2}\sum_{\ell'_3=1}^{3} G(\ell_1\ell_2\ell_3\ell'_1\ell'_2\ell'_3) \times u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell'_1 j'} u_{\ell'_2 k'} u_{\ell'_3 m'} \quad (37)$$

を得る．α の値に拘わらず，$\ell_1 = \ell_3 = 1$ 及び $\ell_2 = 2$ が条件を満たすという点は変わらなかった．そこで i を順番に

表 2 選択された 163 遺伝子と SARS-CoV-2 のタンパクと相互作用することが知られているヒトタンパクとの一致度（フィッシャーの正確検定によるP値とオッズ比．詳細は先行研究参照 [3].

Table 2 Coincidence between 163 genes and human proteins that are known to interact with SARS-CoV-2 proteins during infection. The *P*-values were computed by applying Fisher's exact tests to the confusion matrix. For details, refer to [3].

SARS-CoV-2 proteins	P values				Odds Ratio			
		RBF kernel				RBF kernel		
	先行研究 [3]	$\alpha = 10^{-2}$	10^{-4}	10^{-6}	先行研究 [3]	$\alpha = 10^{-2}$	10^{-4}	10^{-6}
SARS-CoV2 E	6.55×10^{-27}	1.08×10^{-44}	4.90×10^{-22}	6.58×10^{-26}	10.16	16.18	8.63	9.84
SARS-CoV2 M	1.37×10^{-26}	1.66×10^{-37}	4.14×10^{-22}	4.14×10^{-22}	8.41	11.40	7.26	7.26
SARS-CoV2 N	4.61×10^{-24}	7.22×10^{-38}	3.51×10^{-15}	1.46×10^{-29}	11.42	17.11	7.93	13.63
SARS-CoV2 nsp1	1.05×10^{-20}	3.98×10^{-28}	4.26×10^{-14}	5.51×10^{-15}	10.00	12.92	7.43	7.78
SARS-CoV2 nsp10	3.40×10^{-20}	9.04×10^{-26}	4.44×10^{-17}	3.88×10^{-19}	11.51	14.13	10.06	11.02
SARS-CoV2 nsp11	1.13×10^{-29}	2.62×10^{-38}	4.00×10^{-20}	1.2×10^{-27}	10.66	13.43	7.74	10.03
SARS-CoV2 nsp12	4.87×10^{-20}	1.64×10^{-29}	1.85×10^{-14}	2.41×10^{-15}	9.48	13.09	7.38	7.72
SARS-CoV2 nsp13	6.04×10^{-33}	3.08×10^{-37}	2.19×10^{-20}	7.41×10^{-28}	11.17	12.49	7.51	9.65
SARS-CoV2 nsp14	1.75×10^{-22}	1.08×10^{-31}	2.72×10^{-16}	2.78×10^{-18}	12.05	16.26	9.30	10.18
SARS-CoV2 nsp15	1.85×10^{-20}	4.26×10^{-29}	1.71×10^{-17}	1.08×10^{-14}	10.23	13.75	9.03	7.90
SARS-CoV2 nsp2	4.81×10^{-33}	5.16×10^{-42}	3.31×10^{-20}	6.34×10^{-31}	11.79	14.76	7.83	11.11
SARS-CoV2 nsp4	5.79×10^{-29}	5.25×10^{-42}	4.09×10^{-16}	5.53×10^{-26}	10.26	14.41	6.47	9.36
SARS-CoV2 nsp5	3.78×10^{-25}	4.18×10^{-38}	6.97×10^{-18}	4.63×10^{-24}	12.36	17.97	9.36	11.91
SARS-CoV2 nsp5^{+}	3.75×10^{-17}	1.42×10^{-21}	3.74×10^{-17}	4.65×10^{-14}	11.39	13.78	11.39	9.72
SARS-CoV2 nsp6	9.47×10^{-26}	7.15×10^{-35}	3.91×10^{-21}	4.97×10^{-22}	9.00	11.71	7.68	7.94
SARS-CoV2 nsp7	1.93×10^{-29}	1.09×10^{-33}	2.70×10^{-16}	1.61×10^{-22}	10.81	12.18	6.76	8.65
SARS-CoV2 nsp8	1.11×10^{-29}	1.68×10^{-41}	1.57×10^{-18}	1.11×10^{-28}	10.14	13.80	6.95	9.84
SARS-CoV2 nsp9	5.36×10^{-29}	1.46×10^{-40}	8.17×10^{-19}	6.71×10^{-28}	12.24	16.67	8.59	11.85
SARS-CoV2 orf10	5.29×10^{-34}	2.44×10^{-44}	5.57×10^{-20}	1.16×10^{-26}	12.37	15.95	7.94	10.01
SARS-CoV2 orf3a	2.06×10^{-28}	1.31×10^{-43}	3.17×10^{-18}	1.62×10^{-24}	9.95	14.76	6.99	8.80
SARS-CoV2 orf3b	1.89×10^{-29}	7.58×10^{-41}	1.35×10^{-17}	2.50×10^{-27}	11.80	15.94	7.78	11.06
SARS-CoV2 orf6	8.81×10^{-26}	5.67×10^{-39}	5.22×10^{-13}	8.03×10^{-22}	10.37	14.99	6.14	9.05
SARS-CoV2 orf7a	1.69×10^{-28}	9.79×10^{-35}	8.26×10^{-22}	1.18×10^{-23}	10.00	11.90	8.04	8.57
SARS-CoV2 orf8	5.94×10^{-28}	2.48×10^{-39}	4.32×10^{-18}	1.27×10^{-20}	9.25	12.57	6.57	7.25
SARS-CoV2 orf9b	6.54×10^{-30}	7.07×10^{-44}	5.12×10^{-17}	9.11×10^{-28}	12.12	17.34	7.68	11.36
SARS-CoV2 orf9c	1.11×10^{-28}	1.03×10^{-42}	1.15×10^{-21}	5.76×10^{-26}	8.35	12.07	6.66	7.69
SARS-CoV2 Spike	8.22×10^{-26}	3.48×10^{-41}	8.46×10^{-15}	6.07×10^{-22}	10.08	15.34	6.55	8.81

+ with mutation, C145A

除いて $u_{1j}u_{2k}u_{1m}$ （つまり $\ell_1 = \ell_3 = 1, \ell_2 = 2$）を計算して k の値との相関係数の変化を計算した．より大きく相関係数の絶対値が劣化する順に「テンソル分解を用いた教師なし学習による変数選択法」で選ばれたのと同じ１６３遺伝子を選ぶことで遺伝子選択を行った．その結果が表 2 に書かれている．この結果 $\alpha = 0.01$ の場合に提案手法は「先行研究」すなわち「テンソル分解を用いた教師なし学習による変数選択法」を超える性能をだせることが分かった．やはり *large p small n* 問題で提案手法が従来の「テンソル分解を用いた教師なし学習による変数選択法」を越えることができるかどうかは具体的な問題をやってみなくては分からないのである．

3.4 腎臓がんの診断遺伝子

すでに紙面が尽きているので詳細は述べられないが，この他にも原著論文 [2] では，腎臓がんの診断遺伝子の探索に「テンソル分解を用いた教師なし学習による変数選択法」を用いた研究 [4] と提案手法の比較を行った．「テンソル分解を用いた教師なし学習による変数選択法」を用いた研究では $N \times K$ の大きな行列にＳＶＤを適用しなくてはならなかったが，提案手法であれば $M \times M$ の小さいサイズの行列にＳＶＤを適用すれば済んだ．残念ながら提案手法は「テンソル分解を用いた教師なし学習による変数選択法」を用いた研究を越えることができなかったが，圧倒的に小さな行列を扱いながら同等程度の性能を発揮できたことだけは触れておく．

4. 終わりに

本研究報告では原著論文 [2] で紹介した「カーネルテンソル分解を用いた教師なし学習による変数選択法」のあらましについて述べた．この提案手法は当初のターゲットである *large p small n* 問題においても，先行手法である「テンソル分解を用いた教師なし学習による変数選択法」を超える性能を出すことができることが確認された．今後使用するカーネルを工夫することなどにより，より大きな性能向上が得られることが期待される．

謝辞 本研究は科研費 19H05270, 20H04848, 20K12067 の補助を受けて実行された．

参考文献

[1] Y.-h. Taguchi, Unsupervised Feature Extraction Applied to Bioinformatics, Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-22456-1
[2] Y.-h. Taguchi and T. Turki, “Mathematical formulation and application of kernel tensor decomposition based unsupervised feature extraction,” Knowledge-Based Systems, 2021. https://doi.org/10.1016/j.knosys.2021.106834 *in press.*
[3] Y.-h. Taguchi and T. Turki, “A new advanced in silico drug discovery method for novel coronavirus (sars-cov-2) with tensor decomposition-based unsupervised feature extraction,” PLOS ONE, vol.15, no.9, pp.1–16, 09 2020. https://doi.org/10.1371/journal.pone.0238907
[4] K.-L. Ng and Y.-H. Taguchi, “Identification of miRNA signatures for kidney renal clear cell carcinoma using the tensor-decomposition method,” Scientific Reports, vol.10, no.1, 15149 Sept. 2020. https://doi.org/10.1038/s41598-020-71997-6