同時テンソル分解の拡張による医学データの統合

東京医科歯科大学難治疾患研究所　阿部興
名古屋大学大学院医学系研究科／東京医科歯科大学難治疾患研究所　島村徹平

1　はじめに

生命科学の分野においては, 生体内に存在する分子を網羅的に観測することが可能になった. これらは考察される対象の分子（これをモダリティと呼ぶことにする）が遺伝子であればゲノミクス，転写物であればトランスクリプトミクス，代謝物であればメタボロミクスといった様々な分野があり, 総称としてはオミクス（omics）データと呼ばれる. 中でも, ある 1 つのサンプルに対し複数のモダリティを計測するマルチオミクスデータは, 生命現象についての豊富な情報を持つものとして注目されている. 一方で, マルチオミクスデータをどのように分析すべきかは未だ議論が続いている.

複数のモダリティをまたいだ分析を難しくする原因の一つとして, 計測に関わる現実的な制約のために, 対応があるサンプルとないサンプルが混在する semi-paired なデータが多く見られることがあげられる. また, あるモダリティは計数データ, あるモダリティは連続値のデータといった形で, データの分布が大きく変わることも統一的な分析を難しくしている.

Abe & Shimamura (2023) では NMF を拡張し, 多次元のデータを柔軟に分析するモデルとして, unified non-negative matrix factorization (UNMF) を提案した. しかし, UNMF はポアソン分布を仮定していたため, 本報告ではより幅広いクラスのデータを柔軟にあつかえる枠組みについて考察する.

2　モデルと推定アルゴリズム

2.1　概要

確率モデルとしての定式化の前に, まず我々が提案する手法の概要を, そのモチベーションが理解できるように述べたい. 簡単のため, 3 階のテンソル $Y = (y_{i,j,k})$ を考える. ここでは単に添字（i, j, k）を定めたとき値が 1 つに決まるような変数（多次元配列）という意味でテンソルという語を用いる. さらに $(vec)(Y)$ を適当な順番で Y のすべての要素をベクトルに配置する作用素とする. CP 分解（正準分解や並行因子分解とも呼ばれる）では各要素が次を満たすような行列 $V^{(k)}$ $(k = 1, 2, 3)$ を作る.

$$y_{ijk} \approx \sum_r v_{ir}^{(1)} v_{jr}^{(2)} v_{kr}^{(3)}.$$

この式は次のように書き換えられる:

$$y_n \approx \sum_r \prod_{d=1}^{D} v_{dr}^{x_{nd}} \tag{1}$$

ここで

$$V = (v_{dl}) = \begin{pmatrix} v^{(1)} \\ v^{(2)} \\ v^{(3)} \end{pmatrix}$$

and $x_{nd} \in \{0, 1\}$ is one-hot encoded matrix (ξ_1, ξ_2, ξ_3). This representation not conflict the cases that even when the number of subscripts increases or duplicate, even when there are missing values. This fact means our methods can be easily handle semi-paired samples with multi-modal (see Fig. 1).

また交互作用項を考えることは

$$\begin{aligned} y_{ijk} &\approx \sum_r \prod_{d=1}^{D} v_{dr}^{x_{nd}} \\ &= \begin{cases} \sum_r v_{ijr}^{(1)} v_{jr}^{(2)} v_{kr}^{(3)} & \text{when } X \text{ consisiting from } \xi_1 \cdot \xi_2, \xi_2, \text{ and } \xi_3 \\ \sum_r v_{ir}^{(1)} v_{jkr}^{(2)} v_{kr}^{(3)} & \text{when } X \text{ consisiting from } \xi_1, \xi_2 \cdot \xi_3, \text{ and } \xi_3 \end{cases} \end{aligned}$$

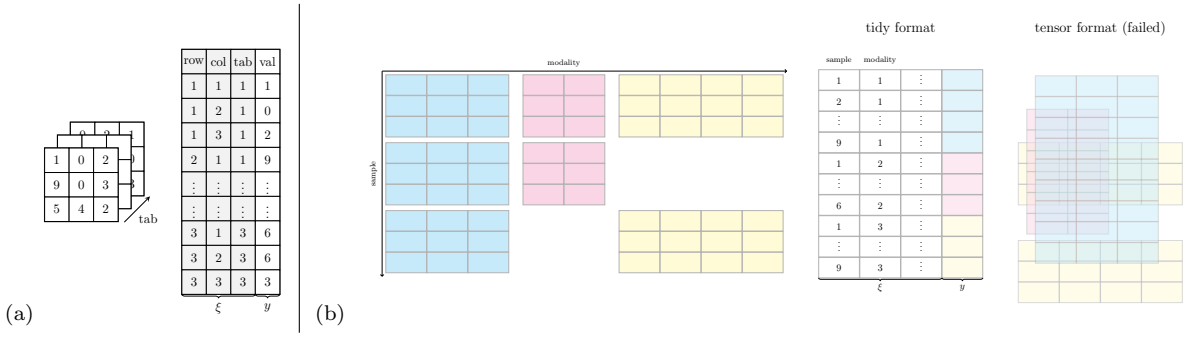

図 1　Conceptual view for semi-pared data integration. (a)Arbitrary m-way tensor can be replanted as tidy-format. The left matrices and the right tidy data structure are equivalent to each other. The all of the index set ξ is map to dummy encoded vectors x_n (b) Data that is difficult to represent as a multidimensional array, can also be stored as tidy-format. White cell indicates a missing value.

このことも semi-paired なデータの分析にとって有益である. 例えば, 行の軸では対応があるが列の軸では対応のない行列を分析するときは, 列ダミーとモダリティの交互作用項を考える.

2.2　モデル

$$y_n \sim \mathcal{N}\left(y_n \mid \sum_{l=1}^{L}\prod_{d=1}^{D} v_{dl}^{x_{nd}}, \lambda^{-1}\right) \tag{2}$$

$$v_{dl} \sim \mathcal{N}(v_{dl}|0,\tau^{-1}) \tag{3}$$

$$\lambda \sim \mathcal{G}(\lambda|a,b)$$

ここで $V=(v_{dl})$ は $D\times L$ 行列,

■Non-negative constraint　For realizing non-negative constraint, use the following truncated normal prior distribution, instead of the normal prior distribution in Eq.3:

$$v_{dl} \sim \mathcal{TN}(v_{dl}|0,\tau^{-1}) \tag{4}$$

where $\mathcal{TN}(x|\mu,\sigma^2)$ is left truncated normal distribution (truncated at 0) with the mean parameter μ and the variance parameter σ^2, respectively. Note that truncated normal density can be written as follows:

$$\mathcal{TN}(x|\mu,\sigma^2) \propto \mathcal{N}(x|\mu,\sigma^2)\mathbb{1}(x)$$

where $\mathbb{1}(x)$ is indicator function as follows:

$$\mathbb{1}(x) = \begin{cases} 1 & x>0 \\ 0 & x\le 0 \end{cases}.$$

In principle, v_dl, which is constrained to be non-negative, can be selected for each subscript d, but in our implementation, only d specified as the first variable allows negative values (`constraint = SN` mode in the package `moltenCP`). This is to avoid complicated interpretations of $(-1)\times(-1)$. Depending on the order on the columns, we can choose a variable which takes a negative value.

■Auxiliary variables When the support of the distribution of y is clearly restricted, we introduce the following auxiliary variable z_n in the model of Eq.2.

$$y_n = A(z_n), \quad z_n \sim \mathcal{N}\left(\sum_{l=1}^{L}\prod_{d=1}^{D} v_{dl}^{x_{nd}}, \lambda^{-1}\right). \tag{5}$$

The function $A(x)$ can be chosen as follows:

- If the observed y_n takes continuous values with no exact zeros, we would use an identity function
$$A(x) = x.$$
- y_n is non-negative:
$$A(x) = \begin{cases} x, & x > 0 \\ 0 & x \leq 0 \end{cases}.$$
- y_n is binary(0 or 1):
$$A(x) = \mathbb{1}(x).$$
- y_n is non-negative integer:
$$A(x) = \begin{cases} \lceil x \rceil, & x > 0 \\ 0 & x \leq 0 \end{cases}.$$

Utilizing the normal density as the observation model implies using a symmetric loss function to evaluate the fit. When the analyst already knows that there are no negative values in the observed data (for example, in the case of weight, length, or count data), predicting 0 as 1 is not the same as predicting it as -1. In these cases, it makes sense to change the support of the model's density function.

When we want to analyze multiple matrices (tensors) simultaneously, the data distribution may vary depending on the modality. We consider this issue in the next.

■Multi-modal analysis Next, we consider that multiple modes and we want to change the distribution for each mode m. The observation model in Eq.2 is switched to following Eq.6

$$y_n = A_{m(n)}(z_n), \quad z_n \sim \mathcal{N}\left(\sum_{l=1}^{L}\prod_{d=1}^{D} v_{dl}^{x_{ndm}}, \lambda_m^{-1}\right) \tag{6}$$

where $m(n)$ is a function which maps index n to mode $m(n)$.

■Offset factor We often want to correct the observed y_n using the known "weight" w_n; such as, the size when the library size is different, or the time when the time required for measurement is different. In such case, we can address this by making a simple change to Eq.6 as following:

$$y_n = A_{m(n)}(z), \quad z_n \sim \mathcal{N}\left(y_n \mid w_n \sum_{l=1}^{L}\prod_{d=1}^{D} v_{dl}^{x_{nd}}, \lambda^{-1}\right). \tag{7}$$

When the all elements of $w = (w_1, \ldots, w_N)$ is 1, this observation model is equivalent to Eq.6.

■Log-joint density In above mentioned settings, the log likelihood $\ell(v_{dl})$ is represented as following:

$$\begin{aligned}\ell(v_{dl}) &= \sum_{n=1}^{N} \log p(z|V, X, w) \\ &= \sum_{n=1}^{N} \left(-\frac{\lambda}{2} \left\{ z_n - \sum_{l=1}^{L} \prod_{d=1}^{D} v_{dl}^{x_{nd}} \right\}^2 \right) + C \\ &= -\sum_{m=1}^{M} \frac{h_{dlm}}{2} \left(v_{dl}^2 - 2 v_{dl} \frac{\eta_{dlm}}{h_{dlm}} \right) + C,\end{aligned}$$

where C is constant term which not depends on v_{dl}, and let η_{dlm} and h_{dlm} be as follows;

$$\eta_{dlm} = \sum_{n} x_{ndm} \prod_{d' \neq d} v_{dl}^{x_{ndm}} \left(z_{nm} - \sum_{l' \neq l} \prod_{d' \neq d} v_{dl}^{x_{ndm}} \right) \tag{8}$$

$$h_{dlm} = \sum_{n} \lambda_m x_{ndm} \prod_{d' \neq d} v_{dl}^{2x_{ndm}}. \tag{9}$$

In the next subsection, we derive the variational EM algorithm from this logarithmic joint density.

2.3 Estimation procedure

2.3.1 Variationl EM algorithm

This subsection we derived estimator of the latent variables based on the variationl EM algorithm[6]. Using mean-field approximation (i.e. suppose that variational posterior distributions $q(v_{dl})$ and $q(\lambda$ are independent) and let $\langle x \rangle$ be the expectation of x under the variational posterior distribution, we get following variational posterior of v_{dl} as following;

$$q(v_{dl}) = \begin{cases} \mathcal{N}(\mu_{dl}, \sigma_{dl}) & \text{if the prior of } v_{dl} \text{ is not truncated} \\ \mathcal{TN}(\mu_{dl}, \sigma_{dl}) & \text{if the prior of } v_{dl} \text{ is truncated,} \end{cases} \tag{10}$$

where v_{dl} and σ_{dl} are defined by

$$\begin{aligned}\mu_{dl} &= \sum_{m=1}^{M} \frac{\langle \eta_{dlm} \rangle}{\langle h_{dlm} \rangle + \tau / \langle \lambda_m \rangle}, \\ \sigma^2 &= \left(\tau + \sum_{m=1}^{M} \langle h_{dlm} \rangle \right)^{-1}.\end{aligned}$$

The variational posterior of λ_m is,

$$q(\lambda) = \mathcal{G} \left(N_m / 2\eta_{dlm}, \left(\sum_{m} h_{dlm} + \tau \right) / 2 \right). \tag{11}$$

The expectations which used for variational updates are listed as following:

$$\begin{aligned}\langle v_{dl} \rangle &= \mu_{dl} + \sigma_{dl} \phi(-\mu_{dl}/\sigma_{dl}) / \Phi(-\mu_{dl}/\sigma_{dl}), \\ \langle v_{dl}^2 \rangle &= \mu_{dl}^2 + \sigma_{dl}^2 + \mu_{dl} \sigma_{dl} \phi(-\mu_{dl}/\sigma_{dl}) / \Phi(-\mu_{dl}/\sigma_{dl}), \\ \langle \lambda_m \rangle &= (N_m \langle \eta_{dlm} \rangle) / \left(\sum_{m} \langle h_{dlm} \rangle + \tau \right).\end{aligned}$$

where $\phi(x)$ and $\Phi(x)$ is standard normal density and distribution function, respectively.

For *local* latent variable z_n, we use Monte-Carlo integral, i.e. sample $\tilde{z}_n$ using variational posterior as follows

- Rectified

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n| - f_n, \sigma_n^2) & y_n = 0, \\ z_n = y_n \text{ with probability } 1 & y_n > 0 \end{cases} \tag{12}$$

- binary

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n| - f_n, \sigma_n^2) & y_n = 0, \\ \mathcal{TN}(z_n|f_n, \sigma_n^2) & y_n = 1 \end{cases} \tag{13}$$

The estimation procedure can be briefly summarized as follows:

- variational E-step: draw $\tilde{z}_n$ for all ns
- variational M-step: update $q(V)$ using Eq.10 and update $q(\lambda)$ using Eq.11

The appendix**??** describes some notes for efficient implementation.

3 シミュレーション

当日はデータ分析事例をあわせて報告する.

参考文献

[1] Abe K & Shimamura T. (2023). UNMF: A unified non-negative matrix factorization for multi-dimensional omics data. *Briefings in Bioinformatics.* (Under Review).