テンソル同時分解の拡張によるオミクスデータの統合

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研究が盛んに行われている. これらの研究は対象となる分子（モダリティ）によって, ゲノム DNA を対象とするゲノミクス，転写物を対象とするトランスクリプトミクス，代謝物を対象とするメタボロミクスなど様々な分野に分類される. これらの分野で得られるデータは総称してオミクス（omics）データと呼ばれる.　特に, 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については, 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 [1]. 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) は非負値行列因子分解（Nonnegative Matrix Factorization; 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 [2]. しかし, UNMF で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提案する.

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, 提案手法の概要を動機となる例を通じて述べる. 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 $\mathrm{vec}(Y)$ 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）ではそれぞれの要素 $v_{il}^{(1)}$, $v_{jl}^{(2)}$, $v_{kl}^{(3)}$ が次を満たすような行列 $V^{(k)}$ $(k = 1, 2, 3)$ を作る.

$$y_{ijk} \approx \sum_l v_{il}^{(1)} v_{jl}^{(2)} v_{kl}^{(3)}.$$

添字 i, j, k のそれぞれの作る集合, ξ_1, ξ_2, ξ_3 をそれぞれワンホットエンコーディングでダミー変数とした行列を X, その (n, d) 成分を x_{nd} とすると, この式は次のように書き換えられる.

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

ここで

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

とした. この表現は図 1 に示す tidy format [3] の利便性と対応しており, 添え字の数が増えた場合, または重複や欠損値が発生する場合でも一貫して扱える. 図 1 中の ξ で示した変数はすべてダミー変数化して計画行列 X を構成する.

また x_{nd} において, モダリティとの 2 次の交互作用項を考えることはモダリティごとに変化することを許すことと同値である. 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, 次のように

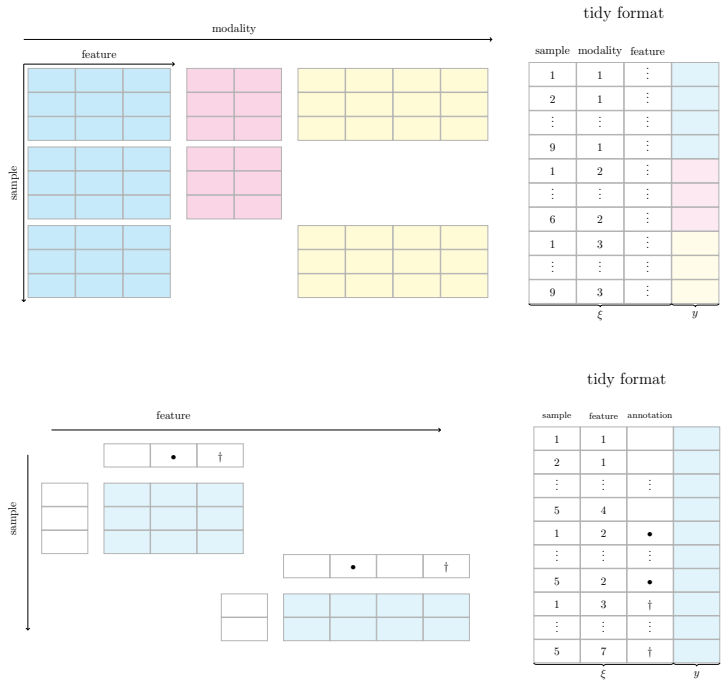

図 1　semi-pared なデータの統合. 多次元配列は右に示す形式（tidy-format）で表現できる. また多次元配列の形式での対応づけが難しいデータも tidy-format で表現できる. （上）複数のモダリティからなるデータ. （下）サンプルごとの対応はないが, 関連しあう特徴量についての情報を持つデータ.

列ダミーとモダリティの交互作用項を考える.

$$y_{ijk} \approx \sum_{l} \prod_{d=1}^{D} v_{dl}^{x_{nd}} = \sum_{l} v_{il}^{(1)} v_{jkl}^{(2)} v_{kl}^{(3)} \quad \text{when } X \text{ consisiting from } \xi_1,\ \xi_2 \cdot \xi_3, \text{ and } \xi_3$$

実際にはこの他にも様々な x_{nd} を考え得る. このこともまた semi-paired なデータの分析にとって有益である. 次の節でここまでの議論を確率モデルとして定式化する.

2.2　モデル

まずいくつか記号を定める. 大きさ N の組の変数 (y_n, ξ_n), $(n = 1, \ldots, N)$ が観測されるとする. ξ_n は y_n と対応する任意の質的変数のベクトルである. また $y = (y_1, \ldots, y_N)'$ とまとめて表記する. ξ_n は（例えばワンホットエンコーディングを用いることによって）2 値変数のベクトル x_n に写すことができる. この記法の下で, 次のデータ生成過程を考える.

$$\begin{aligned} y_n &= A_{m(n)}(z_n), \quad (z_n \mid V, \lambda) \sim \mathcal{N}\left(z_n \mid \sum_{l=1}^{L} \prod_{d=1}^{D} v_{dl}^{x_{ndm}}, \lambda_m^{-1}\right) \\ \lambda &\sim \mathcal{G}(\lambda | a, b). \end{aligned} \tag{2}$$

ここで $\mathcal{N}(x|\mu, \sigma^2)$ は平均と分散がそれぞれ μ と σ^2 の正規分布を表し, $\mathcal{G}(x|a, b)$ は形状パラメータとレートパラメータがそれぞれ a と b のガンマ分布を表す. $V = (v_{dl})$ は $D \times L$ 行列であり推定の対象となる潜在変数である. また, z_n, $A_m(z)$, $m(n)$ という新しい記号を導入したので順に説明する. これらはモダリティによって y の取りうる値の範囲が明確に変わる場合を考えるためのものである. 例を上げると, データに負の値が生じないことが明らかなとき, 0 を 1 と予測することと -1 と予測することは対称に評価できない. そのため, 分布の台の変化に対応すべく, 中間的な変数 z_n を考え, 関数 $A_m(x)$ を次のいずれかから分析者が選択する.

- y_m が実数（連続）値をとるとき：
$$A_m(x) = x.$$
- y_m が非負のとき:
$$A_m(x) = \begin{cases} x, & x > 0 \\ 0 & x \leq 0 \end{cases}.$$
- y_m が 2 値 (0 or 1) のとき:
$$A_m(x) = \mathbb{1}_{(0,\infty)}(x).$$
- y_m が非負の整数のとき:
$$A_m(x) = \begin{cases} \lfloor x \rfloor, & x > 0 \\ 0 & x \leq 0 \end{cases}.$$

ここで $\mathbb{1}_A(x)$ は $x \in A$ が成り立つとき 1, 他の場合 0 の値を取る指示関数とした.

$m(n)$ は n をモダリティを区別する添字に写す関数とした. これにより複数のモダリティを考慮し, 各モダリティの分布を変更できる.

また v_{dl} の事前分布については, 次のように非負制約の有無によって使い分ける.

$$v_{dl} \sim \mathcal{N}(v_{dl}|0, \tau^{-1}), \quad \text{（非負のとき）}, \tag{3}$$
$$v_{dl} \sim \mathcal{TN}(v_{dl}|0, \tau^{-1}), \quad \text{（負の値を許すとき）}. \tag{4}$$

ここで $\mathcal{TN}(x|\mu, \sigma^2)$ は 0 で左側を切断された正規分布である. この切断正規分布は次のように書ける.

$$\mathcal{TN}(x|\mu, \sigma^2) \propto \mathcal{N}(x|\mu, \sigma^2)\mathbb{1}_{[0,\infty)}(x)$$

非負制約の有無は v_{dl} ごとに設定できるとしてもこの節の議論に矛盾は生じないが, 我々の実装では指定された 1 つの変数のみに負の値を許すことにした. これは $1 = (-1) \times (-1)$ によって解釈が煩雑になるのを避けるためである.

ここまで述べたモデルに基づき v_{dl} についての対数尤度関数 $\ell(v_{dl})$ は次のように書ける.

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

ここでは v_{dl} に依存しない項を C とまとめて置き, η_{dl} と h_{dl} は次のように置いた.

$$\eta_{dl} = \sum_n x_{nd} \prod_{d' \neq d} v_{dl}^{x_{ndm}} \left(z_n - \sum_{l' \neq l} \prod_{d' \neq d} v_{dl}^{x_{nd}} \right), \quad h_{dl} = \sum_n \lambda x_n \prod_{d' \neq d} v_{dl}^{2x_{nd}}.$$

ここから推定のためのアルゴリズムを導くことができる.

2.3 変分ベイズ法

この節では変分 EM アルゴリズム [4] に基づき, 提案モデルについての推定量を導出する. 本研究では平均場近似の仮定を置く. すなわち, 近似事後分布が互いに独立とした下でカルバックライブラ情報量の意味で事後分布を近似する $q(v_{dl})$ と $q(\lambda)$ を求める. また変分事後分布による x の期待値を $\langle x \rangle$ と表記する. v_{dl} についての変分事後分布として次を得る.

$$q(v_{dl}) = \begin{cases} \mathcal{N}(\mu_{dl}, \sigma_{dl}) & \text{when the prior of } v_{dl} \text{ is not truncated} \\ \mathcal{TN}(\mu_{dl}, \sigma_{dl}) & \text{when the prior of } v_{dl} \text{ is truncated,} \end{cases} \tag{5}$$

ここで μ_{dl} と σ_{dl} は次のように置いた.

$$\mu_{dl} = \frac{\langle \eta_{dl} \rangle}{\langle h_{dl} \rangle + \tau / \langle \lambda \rangle}, \quad \sigma^2 = (\tau + \langle h_{dl} \rangle)^{-1}.$$

λ についての変分事後分布として次を得る.

$$q(\lambda) = \mathcal{G}\left(N/2\eta_{dl}, \left(\sum_m h_{dl} + \tau \right) /2 \right). \tag{6}$$

変分ベイズ法の更新式で必要な各確率変数の期待値は次の通りである.

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

ここで $\phi(x)$ と $\Phi(x)$ をそれぞれ標準正規分布の確率密度関数, 分布関数とした.

潜在変数 z_n, についての期待値は Monte-Carlo 積分で評価する. z_n の Monte-Carlo サンプルを特に $\tilde{z}_n$ と書くことにし, 次の変分事後分布から疑似乱数でサンプルする.

- y_n が実数（連続）値をとるとき：

$$z_n = y_n \quad \text{(with probability 1)} \tag{7}$$

- y_n が非負のとき:

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n| - f_n, \sigma_n^2), & y_n = 0, \\ z_n = y_n \text{ with probability 1}, & y_n > 0. \end{cases} \tag{8}$$

- y_n が 2 値 (0 or 1) のとき:

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n| - f_n, \sigma_n^2), & y_n = 0, \\ \mathcal{N}(z_n|f_n, \sigma_n^2), & y_n = 1. \end{cases} \tag{9}$$

- y_n が非負の整数のとき:

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n| - f_n, \sigma_n^2), & y_n = 0, \\ \mathcal{N}(z_n|f_n, \sigma_n^2)\mathbb{1}_{[y_n, y_n+1]}(z_n) \cdot C, & y_n > 0 \end{cases} \tag{10}$$

ここで C は z_n に依存しない定数とした.

要約すると, 近似事後分布を実現するアルゴリズムは次のようになる：

- 変分 E ステップ: 式 7–10 を用いて $\tilde{z}_n$ をサンプルする.
- 変分 M ステップ: 式 5 を用いて $q(v_{dl})$ を, 式 6 を用いて $q(\lambda)$ を更新する.

当日はこの推定量の性質を評価するシミュレーションと, 実際的なデータ分析事例をあわせて報告する.

参考文献

[1] Heumos L, Schaar AC, Lance C, et al. (2023). Best practices for single cell analysis across modalities. *Nature Review Genetics*, 24, 550-572.

[2] Abe K & Shimamura T. (2023). UNMF: A unified non-negative matrix factorization for multi-dimensional omics data. *Briefings in Bioinformatics*. 24 (5), bbad253.

[3] Wickham H. (2019). *Advanced R, second edition*. Chapman and Hall/CRC.

[4] Jordan MI, Ghahramani Z, Jaakkola TS & Saul LK. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2), 183-233.