非負値行列因子分解の統一的な拡張と医学データ解析

名古屋大学大学院医学系研究科　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, 次世代シークエンシングに代表される実験技術の進歩に伴い, ゲノム, エピゲノム, トランスクリプトーム, プロテオーム, メタボロームといった, 多数の要素を縦断的に観測することが可能になった. これらのデータは, 非負の整数（カウント）データである事が多く, 計算機上では多次元配列（テンソルと呼ばれることもある）を使用して表すことができる. そして, 環境因子と遺伝子型, 表現型との関連を明らかにするための豊富な情報を持つ. 一方で, データが高次元になるため, 多変量のデータの次元削減やパターンの抽出を行う分析手法の開発が求められている.

多変量解析はかねてより研究されてきたテーマである.　主成分分析（Tipping & Bishop, 1999）, 潜在的ディリクレ配分（Blei & Jordan, 2003）, 非負値行列因子分解（NMF; Cemgil, 2009）といった, データの次元削減やパターンの抽出を行うことを目的とした統計モデルの多くは, 行列分解を行う手法と捉えることができる.　テンソル分解は行列分解をさらに多次元へ拡張する方法の一つである. しかし, データを 3 次元以上の多次元配列に格納する場合, 繰り返し測定や補助的な情報がある場合にその都度データの形式を作り変える必要がある. また, 欠損値がある場合には, 欠損をなんらかの方法で補完する, または欠損した部分をマスクするといった処理が必要になる.

これに対して, tidy data と呼ばれる次の条件を満たす形式でデータを取り扱うことが提唱されている (Wickham, 2019).

1. 各変数が 1 つの列を形成する
2. 各観測値が 1 つの行を形成する
3. 各観測の観測単位の種類が 1 つのテーブルを形成する
4. 各値が 1 つのセルを構成する

tidy data の形式でデータを処理する場合, 繰り返し測定はテーブルへの行の追加, 欠損はテーブルからの行の削除であり, 大規模な変更は起こらない. さらに, 実験時の条件や補助的な情報を列として, あわせて分析することができる.

Abe & Shimamura (2023) では NMF を tidy data の枠組みに拡張し, 多次元のデータを柔軟に分析するモデル, unified non-negative matrix factorization (UNMF) を提案した. 本報告ではこのモデルを活用し, 数値実験とデータ分析の結果を示す.

2　モデルと推定アルゴリズム

UNMF は, 因子分解と同様に, 潜在因子と観測された変数間の関係を記述する.　観測された $y = (y_1, \ldots, y_N)'$ および計画行列 $X = (x_{nd})$ $(x_{nd} \in \{0, 1\})$ に対し, UNMF は次のような $D \times L$ の非負値行列 $V = (v_{dl})$ を推定する.

$$y_n \approx \sum_{l=1}^{L} \prod_{d=1}^{D} v_{dl}^{x_{nd}}.$$

ここで, D は X の次元（列の数）によって定まり, L はユーザーが指定する. 行列 V は y_n の共起関係を記述するパラメータである. 右辺は X の非ゼロ成分で指定される V の行ベクトルのなす方向どうしが近いときより大きく, 遠いときにはより小さい値となる.

UNMF を使用した tidy data 形式での行列分解を, 簡単な例を通じて紹介する (図 1). 2 人の被験者に対し実験で 3 つの項目の観測値が得られたとする. 実験の結果得られたデータは行で被験者, 列で項目を区別して, 行列 Y で表すことができる. 一方, 実験の結果得られたデータをベクトル $y = \mathrm{vec}(Y)$ に対し, 行列 X を被験者と項目を表すダミー変数で表現することもできる. これは tidy data の形式と直接対応する.　行列で表現

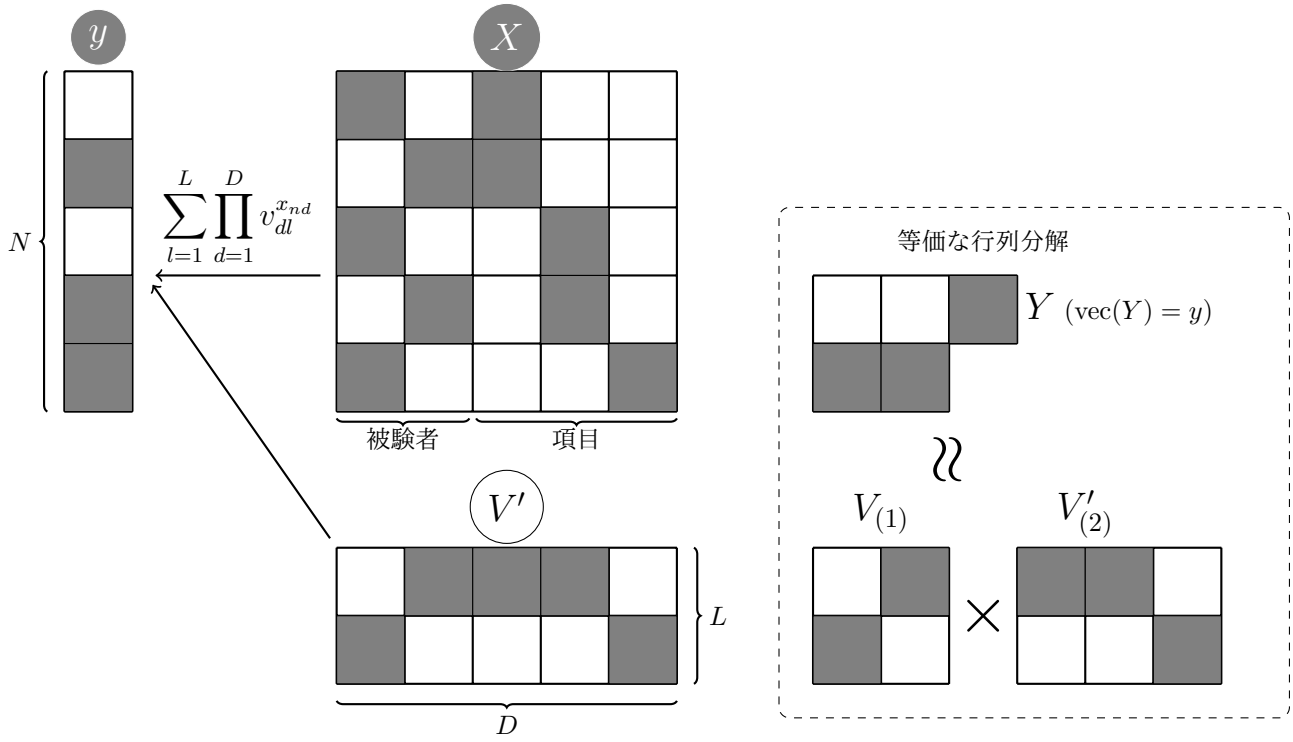

図 1　UNMF の視覚的な例. 灰色の円は観測された変数を示し, 白は潜在変数 .

したときと tidy data の形式で表現したときとで, データの持つ情報は等しい. UNMF は次の式が成り立つように推定を行う.

$$y_n \approx \sum_{l=1}^{L}\prod_{d=1}^{D} v_{dl}^{x_{nd}} = V_{(1)}V'_{(2)}$$

右辺より, この場合 UNMF は NMF と等価である. ここで $V_{(1)} = (\prod_{d=1}^{2} v_{dl}^{x_{nd}})$, $(V_{(2)} = \prod_{d=3}^{5} v_{dl}^{x_{nd}})$　とした.

図 1 では Y の (2,3) 成分が欠損した状況を示している. この場合も UNMF の枠組みではデータに対し特殊な処理は必要とされない. また, 同一条件の重複（繰り返し測定）がある場合も同様である. さらに, 被験者と項目のみならず, 観測した時点やその他の説明変数といった, 豊富な情報を含むデータが得られている場合を含め, 一貫した枠組みで分析を行うことができる.

さて, 具体的な確率モデルとして, UNMF は次のデータ生成過程を考える.

$$\begin{aligned} y_n \mid v_{dl}, x_n &\sim \text{Poisson}\left(\sum_{l=1}^{L}\left\{\prod_{d=1}^{D} v_{dl}^{x_{nd}}\right\}\right), \\ v_{dl} &\sim \text{Gamma}(a, b). \end{aligned} \tag{1}$$

ここで Poisson(λ) は平均 λ のポアソン分布, Gamma(a, b) は形状パラメータ a, レートパラメータ b のガンマ分布, およびそれらの確率密度を意味する. ガンマ分布はパラメータ V に対する事前分布として置いた.

式 (1) は次の表現と等価である.

$$y_n = \sum_{d=1}^{D}\sum_{l=1}^{L} u_{nl}, \quad u_{nl} \sim \text{Poisson}\left(\prod_{d=1}^{D} v_{dl}^{x_{nd}}\right)$$

u_{nl} は推定アルゴリズムを導くための補助的な潜在変数である. 平均場近似の仮定, すなわち推定の対象となる

変数の事後分布がすべて互いの独立であると仮定すると, いくつかの計算から, 事後分布の近似 $q(u_{nl})$ および $q(v_{dl})$ を得ることができる (Jordan, *et al.*, 1999). すなわち, 潜在変数が互いに独立だとした範囲内で, カルバック・ライブラ距離の意味で事後分布に近い分布を求めるという, 変分ベイズ法による推定を行う. 表記として, ある確率変数 w に対し, その近似事後分布 $q(w)$ で評価した平均を $\widetilde{w}$ とする. また, 一般には $\log \widetilde{w}$ と $\widetilde{\log w}$ は一致しないことに注意する. $q(u_{nl})$ は次に示すパラメータ r_{nl} を持つ多項分布,

$$r_{nl} = \frac{\exp(\widetilde{\log v_{dl}})}{\sum_{l=1}^{L} \exp(x_{nd}\widetilde{\log v_{dl}})}, \tag{2}$$

$q(v_{dl})$ は次に示す形状パラメータ $\hat{a}_{dl}$, レートパラメータ $\hat{b}_{dl}$ のガンマ分布である.

$$\begin{aligned} \hat{a}_{dl} &= \sum_{n=1}^{N}\sum_{n=1}^{N} x_{nd}\widetilde{u_{nl}} + a, \\ \hat{b}_{dl} &= \sum_{n=1}^{N} x_{nd}\left(\prod_{d'\neq d} \widetilde{v_{d'l}}^{x_{nd'}}\right) + b. \end{aligned} \tag{3}$$

推定アルゴリズムは次の手順としてまとめられる.

- v_{dl} をガンマ分布で初期化する
- 指定された反復回数まで次のステップを繰り返す
  - 式 (2) を用いて $\widetilde{u_{nl}}$ を更新
  - 式 (3) を用いて $\widetilde{v_{dl}}$ と $\widetilde{\log v_{dl}}$ を更新

3 分析対象

本研究ではシークエンシングと呼ばれる技術を利用して得られたデータを主な分析の対象とする. これらはサンプルから得られる塩基配列を化学的に増幅して標識し, 塩基を特定する. 結果, 得られるのは塩基配列ごとのカウントデータである. この技術は, DNA の塩基配列を元に, 合成された RNA から組織ごとに発現している遺伝子を特定するトランスクリプトーム解析, ある環境中から採取したサンプルにどのような種類の微生物が存在するかを調査するメタゲノム解析といった, 幅広い応用範囲を持つ.

本稿では一例としてメタゲノム解析について記載する. David *et al.* (2014) は、A と B の 2 人のドナーの糞便から採取したメタゲノムを縦断的に調査した. ドナー B については, 530 種の細菌（不明な種は “unknown” としてまとめて集計）が 176 の時点で記録されており, ドナー B は 151 日目から 159 日目の間に腸管感染症に罹患していたことがわかっている.

我々は時間と菌種をワンホットエンコーディングでダミー変数とし, UNMF を用いてこのデータを分析した. データをホールドアウトしたバリデーションにより, $L = 5$ と定めた. 推定された V の近似事後分布による平均について, 図 2 に示す. 図 2 が示すように, 150 から 160 日目にかけて, 成分 4 は顕著に大きい値を取る. この成分 4 は, 食中毒の原因となることが知られている *Eshtishia/Shigella*, *Salmonella*, および *Clostridium* 属によって特徴付けられる. このことは, UNMF によって背景知識と一致する特徴が抽出されたことを意味する. 加えて, 成分 4 と入れ替わるように成分 3 が増加し, 腸管感染症以前に多くの割合を占めていた他の成分は減少した傾向が見て取れる. UNMF はこのように腸内細菌叢の変化を捉えることができる.

大会ではこれらのデータ分析事例に合わせて数値実験の結果を報告する.

参考文献

[1] Tipping ME., & Bishop, CM. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 611-622.

[2] Blei DM, Ng AY, & Jordan MI. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

[3] Cemgil AT. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, 2009. https://doi.org/10.1155/2009/785152

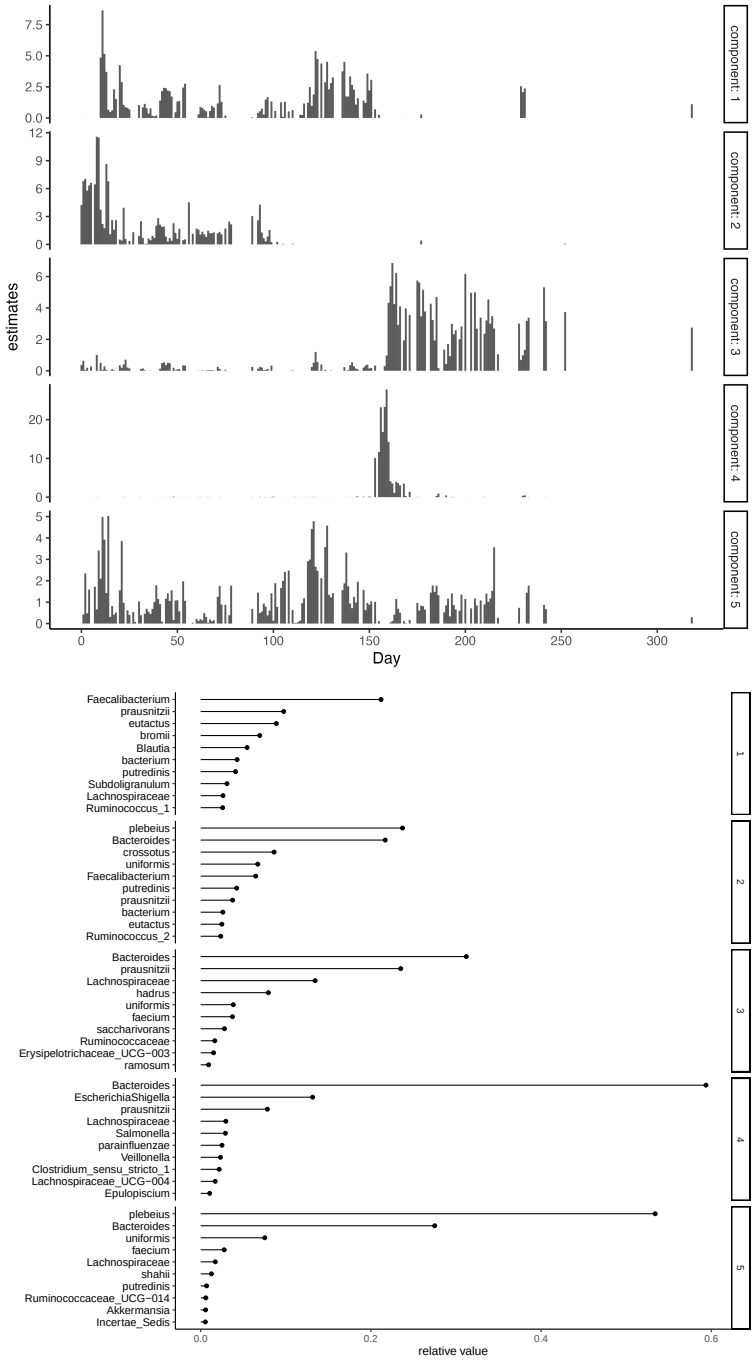

図 2　推定された V. 上段：時間ダミー, 下段：菌種ダミーに対応する. 下段では代表的な細菌上位 10 種のみを示した.

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

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

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

[7] David LA, Materna AC, Friedman J, *et al.* (2014). Host lifestyle affects humanmicrobiota on daily timescales. *Genome Biology*, 15(R89). `https://doi.org/10.1186/gb2014-157-r89`.