

テンソル同時分解の拡張による オミクスデータの統合

阿部興¹・島村徹平²

2023 年 6 月 3 日

¹東京医科歯科大学難治疾患研究所

²名古屋大学医学系研究科・東京医科歯科大学難治疾患研究所

動機：分析対象

モダリティ

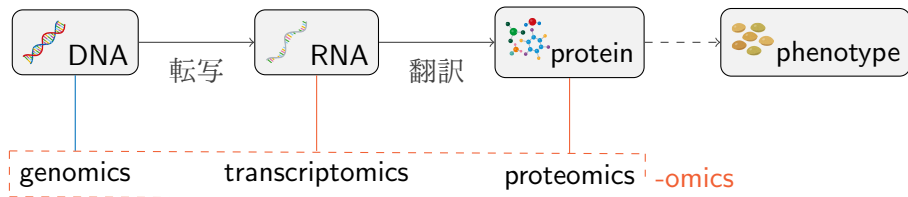


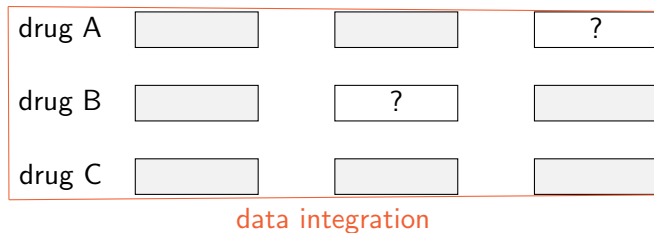
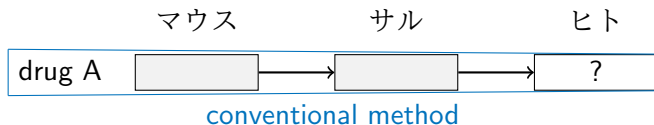
Fig: 遺伝子発現の情報の流れ

オミクス (omics) データを統合して分析したい

積極的理由：データを補い合い普遍的な特徴を抽出

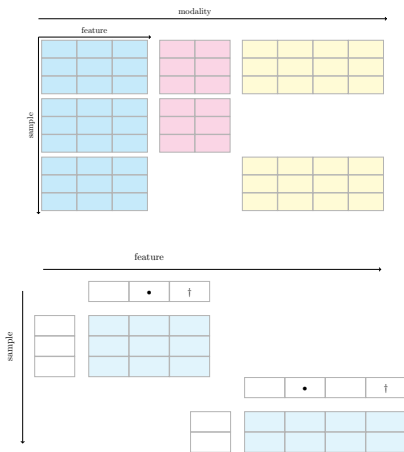
消極的理由：対応のあるサンプルなので非独立

課題：データ統合



- ▶ semi-paired なデータが多い
- ▶ モダリティごとに分布が変わる

multi-dimensional array



tidy format

sample	modality	feature	
1	1	⋮	
2	1	⋮	
⋮	⋮	⋮	
9	1	⋮	
1	2	⋮	
⋮	⋮	⋮	
6	2	⋮	
1	3	⋮	
⋮	⋮	⋮	
9	3	⋮	

sample	feature	annotation	
1	1		
2	1		
⋮	⋮	⋮	
5	4		
1	2	•	
⋮	⋮	⋮	
5	2	•	
1	3	†	
⋮	⋮	⋮	
5	7	†	

tidy-format* の利便性を生かしたまま行列分解ができないか？

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

例: 3 階のテンソルの場合

Data:

$$Y = (y_{ijk}), \quad y = (y_n) = \text{vec}(Y).$$

CP 分解 :

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

提案法[†] :

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

ここで,

$$V = (v_{dl}) = \left(V^{(1)}, V^{(2)}, V^{(3)} \right)'.$$

[†]Ko ABE and Teppei SHIMAMURA (2023) UNMF: A unified non-negative matrix factorization for multi-dimensional omics data. Briefings in Bioinformatics.

内積としての解釈

$$\begin{aligned}
 y_n &\approx \sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}} = v_{11}^{x_{n1}} v_{21}^{x_{n2}} \cdots v_{D1}^{x_{nD}} + \cdots + v_{1L}^{x_{n1}} v_{2L}^{x_{n2}} \cdots v_{DL}^{x_{nD}} \\
 &= \underbrace{\left(v_{d1}^{x_{nd}} \quad \cdots \quad v_{dL}^{x_{nd}} \right)}_{\text{inner product}} \begin{pmatrix} \prod_{d' \neq d} v_{d'1}^{x_{nd'}} \\ \vdots \\ \prod_{d' \neq d} v_{d'L}^{x_{nd'}} \end{pmatrix}
 \end{aligned}$$



inner product of  > inner product of 

Fig: y_n の値が大きいとき V の X で指定される成分どうしの値が似る

アンサンブルとしての解釈

$$\begin{aligned}
 y_n &\approx \sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}} \\
 &= \frac{1}{L} \sum_{l=1}^L \underbrace{L \exp \left(\sum_{d=1}^D x_{nd} \cdot \log v_{dl} \right)}_{\text{log-linear model}}
 \end{aligned}$$



Fig: 複数の対数線形モデルの平均と捉えられる

モデル（準備）

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

$$y_n \mid V, \lambda \sim \mathcal{N} \left(y_n \mid \sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}}, \lambda^{-1} \right) \quad \text{正規分布}$$

$$v_{dl} \sim \mathcal{TN}(v_{dl} \mid 0, \tau^{-1}) \quad 0 \text{ で左側切断された正規分布}$$

$$\lambda \sim \mathcal{G}(\lambda \mid a, b) \quad \text{ガンマ分布}$$

さらにマルチオミクスデータの分析のため：分布を変える

[‡]cf. Abe & Shimamura (2023) はポアソン分布を仮定.

中間変数：分布を変える

- y_n が実数（連続値）：恒等変換

$$A(z_n) = z_n.$$

- y_n が非負：非負化

$$A(z_n) = \begin{cases} z_n, & z_n > 0 \\ 0 & z_n \leq 0 \end{cases}$$

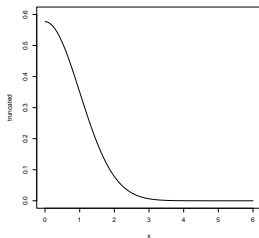
- y_n が 2 値 (0 or 1)：2 値化

$$A(z_n) = \mathbb{1}_{(0, \infty)}(z_n) \quad (\text{indicator function})$$

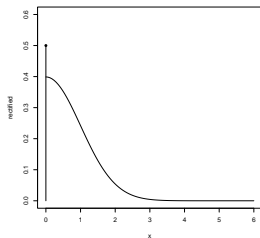
- y_n が非負の整数：離散化

$$A(z_n) = \begin{cases} \lceil z_n \rceil, & z_n > 0 \\ 0 & z_n \leq 0 \end{cases} \quad (\text{ceiling function})$$

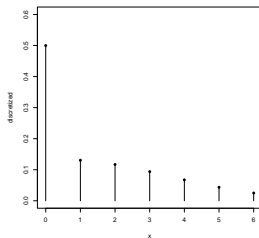
切断, 非負化, 離散化



(a) 切断



(b) 非負化



(c) 離散化

Fig: a: 潜在変数の事前分布. b, c: 観測されるデータのモデル. 観測されるデータのモデルは0の一点で確率を持つ (cf. zero-inflated モデル)

Note: モデルの密度が0になることは対数尤度に $-\infty$ の罰則をつけることと同じであるため, 分布の台が特に重要と考えた.

モデル

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

ここで $m(n)$ は n をモダリティを区別する添字に写す関数.

分析者が設定する要素 (チューニングパラメータ):

- 潜在変数 V の次元 L
- 中間的関数 $A(x)$
- 計画行列 X

対数尤度

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

ここで,

$$\begin{aligned}\eta_{dl} &= \sum_n x_{nd} \prod_{d' \neq d} v_{dl}^{x_{nd}} \left(z_n - \sum_{l' \neq l} \prod_{d' \neq d} v_{dl'}^{x_{nd}} \right), \quad \text{1 次の項} \\ h_{dl} &= \sum_n \lambda x_{nd} \prod_{d' \neq d} v_{dl}^{2x_{nd}}, \quad \text{2 次の項}\end{aligned}$$

v_{dl} に依存しない定数をまとめて C とした.

変分 EM アルゴリズム (一般論)

データ: $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n)$

\mathcal{D} 全体に影響する **global** な潜在変数: V

\mathcal{D}_n に影響する **local** な潜在変数: z_n ($z = (z_1, \dots, z_N)'$)

$$\begin{aligned} D_{KL}(q||p) &= \int q(z) \log \frac{q(z)}{p(z_n|\mathcal{D}, w)} dz \\ &= E_q[\log q(z)] - E_q[\log p(z_n|\mathcal{D}, w)] \end{aligned}$$

$q(z) \propto \exp(E_q[\log p(z_n|\mathcal{D}, w)])$ のとき最小.

変分 EM アルゴリズム

E-step: $q(z_n) \propto \exp(E_q[\log p(z_n|\mathcal{D}, w)])$ を更新

M-step: $q(w) \propto \exp(E_q[\log p(w|\mathcal{D}, z)])$ を更新

local な潜在変数

変分事後分布からサンプリング：

- 非負化：

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n | -\sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}}, \sigma_n^2) & y_n = 0, \\ z_n = y_n \text{ with probability } 1 & y_n > 0 \end{cases}$$

- 2 値化：

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n | -\sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}}, \sigma_n^2) & y_n = 0, \\ \mathcal{TN}(z_n | \sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}}, \sigma_n^2) & y_n = 1 \end{cases}$$

- 離散化：

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n | -\sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}}, \sigma_n^2) & y_n = 0, \\ \mathcal{TN}(z_n | \sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}}, \sigma_n^2) & y_n = 1 \end{cases}$$

global な潜在変数

解析的に期待値計算：

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

ここで,

$$\begin{aligned} \mu_{dl} &= \frac{E_q[\eta_{dl}]}{E_q[h_{dl}] + \tau / E_q[\lambda]}, \\ \sigma^2 &= (\tau + E_q[h_{dl}])^{-1}. \end{aligned}$$

λ の変分事後分布は,

$$q(\lambda) = \mathcal{G}((N/2)E_q[\eta_{dl}], (E_q[h_{dl}] + \tau) / 2).$$

対数周辺尤度の下限

Evidence lower bound (ELBO): 収束の判定とモデル選択に用いる.

$$\begin{aligned}\mathcal{L}(q) &= E_q \left[\log \frac{p(\mathbf{z}, V, \lambda | X, \tau, a, b)}{q(V, \lambda)} \right] \\ &= \left[\sum_n \left(-E_q[\lambda] \sum_n \left\{ \tilde{z}_n^2 + \tilde{z}_n E_q \left[\sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}} \right] \right\} + 0.5 E_q[\log \lambda] \right) \right] \\ &\quad + D_{KL}(q(\lambda) \| p(\lambda)) + \sum_{d,l} D_{KL}(q(v_{dl}) \| p(v_{dl})).\end{aligned}$$

ここでは、乱数でサンプリングする z_n を特に \tilde{z}_n と書いた.

データ分析

- Kostic et al. (2015): 乳幼児の糖尿病に関するヒト腸内細菌叢と代謝物のコホート調査
 - 潜在変数の次元 $L = 3$
 - $A(z)$: 離散化, 非負化
 - X : subject, feature, 日齢, 月齢, Case/Control, modality
- GSE146188 : 5 種の哺乳類 (ヒト, マウス, ブタ, カニクイザル, アカゲザル) の 1 細胞ごとの RNA-seq データ
 - 潜在変数の次元 $L = 10$
 - $A(z)$: 非負化
 - X : cell, gene と種の交互作用, 既知の対応があるヒトの gene

Kostic et al. (2015); 腸内細菌叢と代謝物

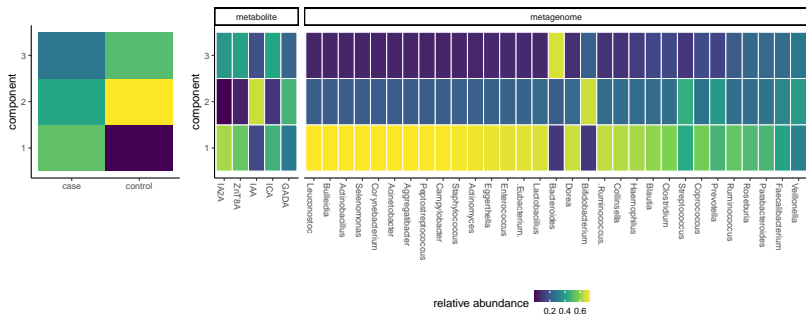
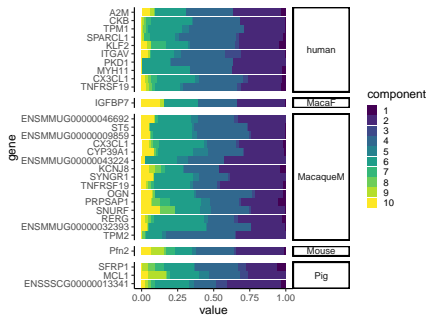


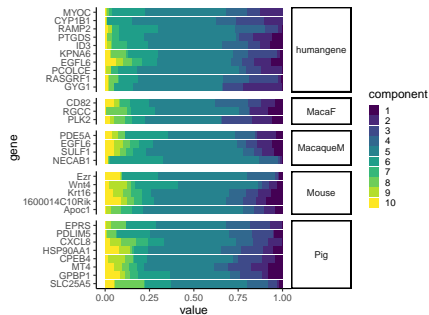
Fig: 推定された V の各 component に関連する細菌・代謝物.

component 1 が相対的に case (糖尿病) に多い. component 2, 3 はそれぞれ Bifidobacterium, Bacteroides が特徴的.

GSE146188; 5 種の哺乳類の RNA-seq



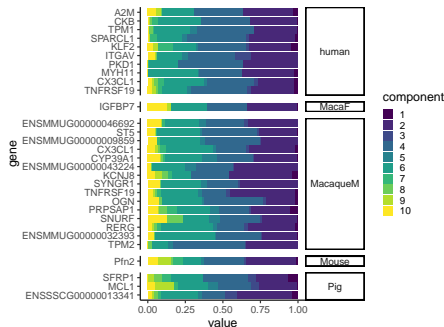
(a) A2M に近い



(b) MYOC に近い

Fig: 推定された V を 2 乗距離で評価

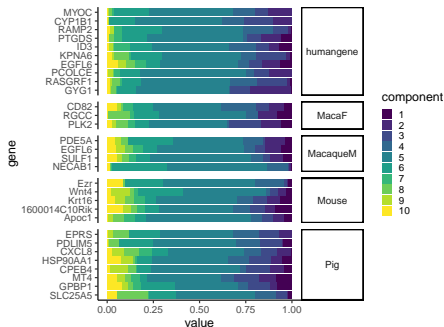
GSE146188; 5 種の哺乳類の RNA-seq



(a) A2M に近い

- A2M: アルツハイマー病と関連
- CKB: 脳内および他の細胞内でホモ二量体として機能
- TPM1: 横紋筋および平滑筋の収縮系および非筋肉細胞の細胞骨格に関与する
- SPARCL1: 解剖学的構造の発達とシナプス組織の調節に関与すると予測される
- KLF2: 哺乳類の発生の初期に発現し、多くの異なる細胞型で見られる

GSE146188; 5 種の哺乳類の RNA-seq



(b) MYOC に近い

- MYOC: 眼圧（緑内障）と関連
- CYP1B1: この遺伝子の変異は、原発性先天性緑内障と関連
- RAMP2: グリコシル化とアドレノメデュリン受容体の細胞表面への輸送に関与
- PTGDS: 平滑筋の収縮/弛緩に関与し、脳で優先的に発現
- ID3: この遺伝子によってコードされるタンパク質は他の HLH タンパク質とヘテロ二量体を形成

まとめ

CP 分解を特殊な場合として含む統計モデルと, 変分ベイズ法に基づくその推定量を提案した.

- 多様なデータに適用できる柔軟性
- 解釈性

今後の発展: 大規模メタアナリシス, 因果推論, 時空間の相関を考慮

`https://github.com/abikoushi/moltenNMF`