

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

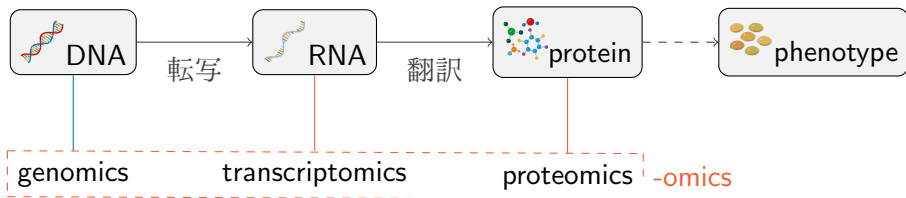
2023 年 6 月 3 日

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



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

動機：データ統合

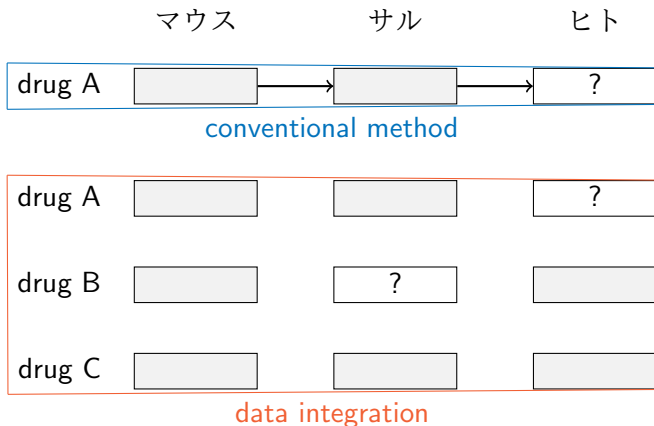
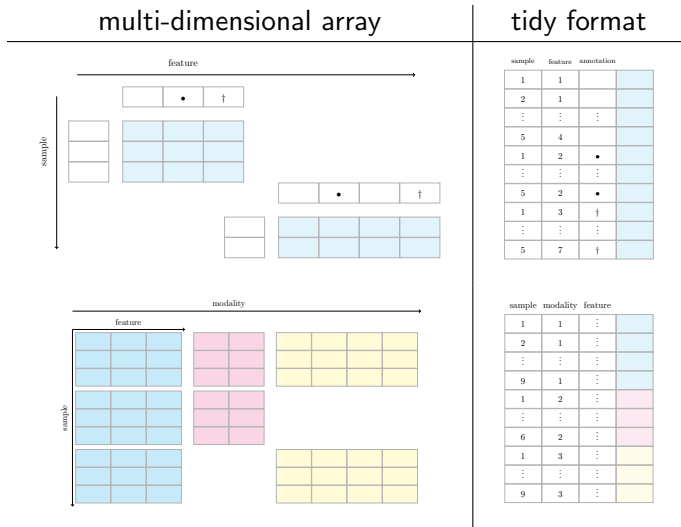


Fig: データ統合により情報を補完しあう

動機：分析手法



tidy-format の利便性を生かしたまま行列分解ができないか？

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 :

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

ここで,

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

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

$$y_n \approx \sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}} = \frac{1}{L} \sum_{l=1}^L \underbrace{L \exp \left(\sum_{d=1}^D x_{nd} \cdot \log v_{dl} \right)}_{f_{nl}}$$

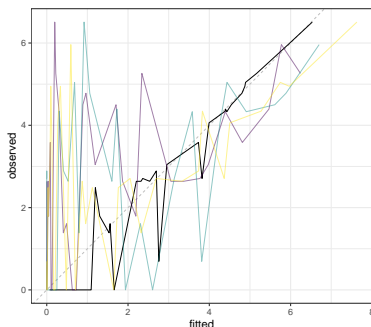


Fig: x 軸: 推定値, y 軸: 実測値, 色付きの線: Lf_{nl} , 黒い線: $\sum_l f_{nl}$

モデル（準備）

式 1 $\left(y_n \approx \sum_l \prod_{d=1}^D v_{dl}^{x_{nd}}\right)$ より

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

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

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

ここで $\mathcal{N}(x \mid \mu, \sigma^2)$ は正規分布（平均 μ , 分散 σ^2 ） $\mathcal{G}(x \mid a, b)$ はガンマ分布（形状パラメータ a , レートパラメータ b ）。

- さらに修正・拡張
 - 解釈性のため：非負制約
 - マルチオミクスデータの分析のため：分布を変える

事前分布：非負制約

$$\begin{aligned} v_{dl} &\sim \mathcal{TN}(v_{dl}|0, \tau^{-1}) \\ &\propto \mathcal{N}(v_{dl}|0, \tau^{-1}) \mathbb{1}_{(0, \infty)}(v_{dl}) \end{aligned}$$

ここで, $\mathbb{1}_A(x)$ は指示関数 ($x \in A$ のとき 1, さもなくば 0) .

Note: 原理的には, 非負制約の有無は変数ごとに選ぶこともできるが, 今回の我々の実装では煩雑さを避けるためすべて非負とした.

中間変数：分布を変える

- y_n が実数（連続値）：

$$A(x) = x.$$

- y_n が非負：

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

- y_n が2値 (0 or 1)：

$$A(x) = \mathbb{1}_{(0,\infty)}(x)$$

- y_n が非負の整数：

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

切断, Rectified, 離散化

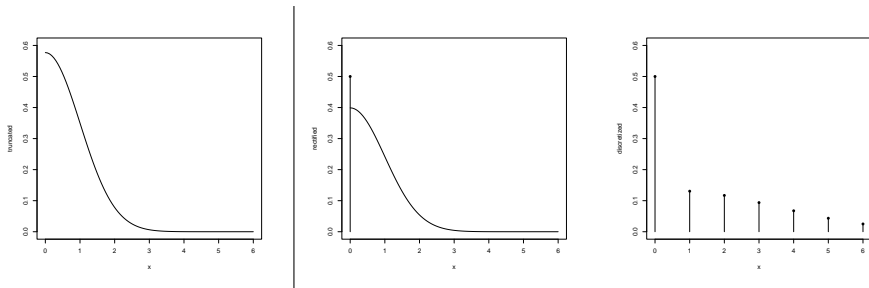


Fig: 左: 潜在変数の事前分布. 右: 観測されるデータのモデル.

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, 分布の台が最も大きい影響を持つと考えた.

モデル

$$y_n = A_{m(n)}(z_n), \quad z_n \sim \mathcal{N} \left(\sum_{l=1}^L \prod_{d=1}^D v_{dl}^{x_{nd}}, \lambda^{-1} \right) \quad (4)$$

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,

$$\eta_{dl} = \sum_n x_{nd} \prod_{d' \neq d} v_{dl}^{x_{nd}} \left(z_n - \sum_{l' \neq l} \prod_{d' \neq d} v_{dl'}^{x_{nd}} \right) \quad (5)$$

$$h_{dl} = \sum_n \lambda x_{nd} \prod_{d' \neq d} v_{dl}^{2x_{nd}}. \quad (6)$$

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

提案モデル; local な潜在変数

変分事後分布からサンプリング：

- rectified

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n | -f_n, \sigma_n^2) & y_n = 0, \\ z_n = y_n \text{ with probability } 1 & y_n > 0 \end{cases} \quad (7)$$

- binary

$$q(-z_n) = \begin{cases} \mathcal{TN}(-z_n | -f_n, \sigma_n^2) & y_n = 0, \\ \mathcal{TN}(z_n | f_n, \sigma_n^2) & y_n = 1 \end{cases} \quad (8)$$

- nonnegative integer

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

提案モデル; global な潜在変数

解析的に期待値計算：

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

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,

$$q(\lambda) = \mathcal{G}((N/2)E_q[\eta_{dl}], (E_q[h_{dl}] + \tau) / 2). \quad (11)$$

データ分析

- Kostic et al. (2015): ヒト腸内細菌叢と代謝物（メタボローム）のコホート調査. 糖尿病予備軍の乳幼児に関する.
- 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, シングルセル（1 細胞ごとの）RNA-seq data.

Kostic et al. (2015)

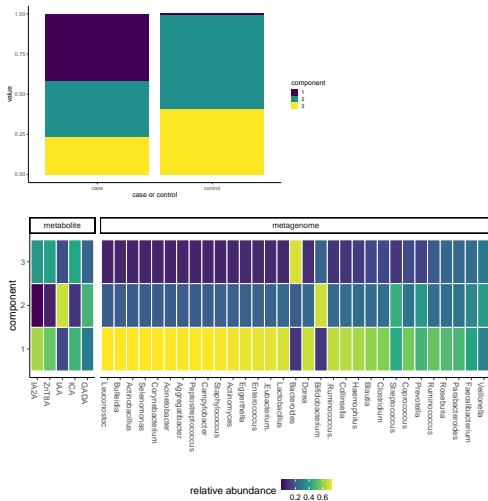


Fig: component 1 が case (糖尿病) に相対的に豊富な細菌のグループとわかる

まとめ

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

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