テンソル同時分解の拡張による オミクスデータの統合

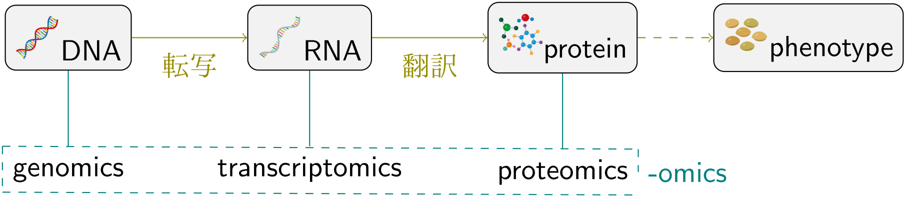
阿部興[1] ・島村徹平[2]

2023 年 6 月 3 日

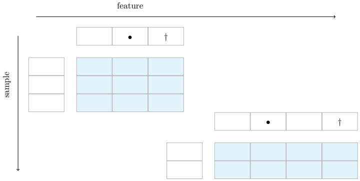
[1]東京医科歯科大学難治疾患研究所
[2]名古屋大学医学系研究科・東京医科歯科大学難治疾患研究所

動機：分析対象

モダリティ

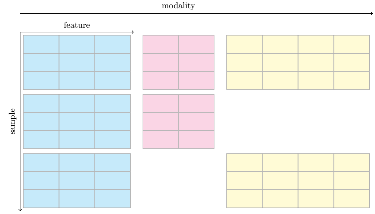


- オミクス（omics）データを統合して分析したい
  - 積極的理由：データを補い合い普遍的な特徴を抽出
  - 消極的理由：対応のあるサンプルなので非独立
- 難しさ
  - semi-paired なデータが多い
  - モダリティごとに分布が変わる



tidy format

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



tidy format

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

tidy-format の利便性をそのままに行列分解ができないか？

3 階のテンソルの場合

Data:

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

CP 分解：

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

提案法：

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

ここで,

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

内積としての解釈

$$\sum_{l=1}^{L}\prod_{d=1}^{D} v_{dl}^{x_{nd}} = v_{11}^{x_{n1}} v_{21}^{x_{n2}} \cdots v_{D1}^{x_{nD}} + \cdots + v_{1L}^{x_{n1}} v_{2L}^{x_{n2}} \cdots v_{DL}^{x_{nD}}$$

$$= \underbrace{\begin{pmatrix} v_{d1}^{x_{nd}} & \dots & v_{dL}^{x_{nd}} \end{pmatrix} \cdot \begin{pmatrix} \prod_{d' \neq d} v_{d'1}^{x_{nd'}} \\ \vdots \\ \prod_{d' \neq d} v_{d'L}^{x_{nd'}} \end{pmatrix}}_{\text{inner product}}$$

アンサンブルとしての解釈

$$
\begin{aligned}
&\sum_{l=1}^{L} \prod_{d=1}^{D} v_{dl}^{x_{nd}} \\
&= \frac{1}{L} \sum_{l=1}^{L} L \underbrace{\exp\left(\sum_{d=1}^{D} x_{nd} \cdot \log v_{dl} \right)}_{\text{let } f_{nl}} \\
&= \frac{1}{L} \sum_{l=1}^{L} L f_{nl} \quad \text{(mean of the log-linear predictor)}
\end{aligned}
$$

モデル（準備）

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

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

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

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

- 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

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

中間変数

- 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

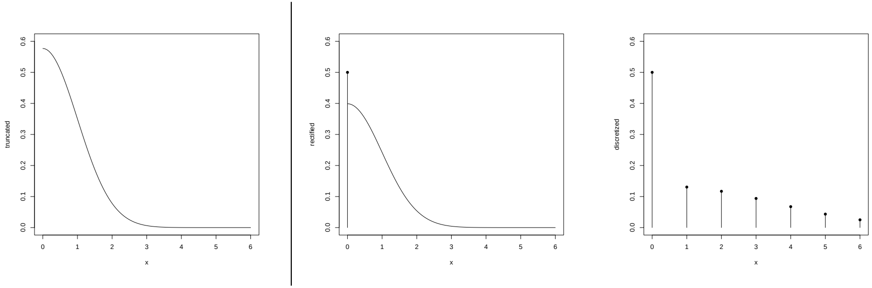


Fig: 左: 潜在変数の事前分布. 右: 観測されるデータのモデル. 一点 0 で確率を持つ（eg. 重さ, 計数）

モデル

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

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

対数尤度

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

ここで,

$$\eta_{dl} = \sum_{n} x_{nd} \prod_{d' \neq d} v_{dl}^{x_{nd}} \left(z_n - \sum_{l' \neq l} \prod_{d' \neq d} v_{dl}^{x_{nd}} \right) \tag{5}$$
$$h_{dl} = \sum_{n} \lambda x_{nd} \prod_{d' \neq d} v_{dl}^{2x_{nd}}. \tag{6}$$

C は v_{dl} に依存しない定数.

変分 EM アルゴリズム（一般論）

global local

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

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

The estimation procedure can be briefly summarized as follows:

- variational E-step: draw $\tilde{z}_n$ for all ns using Eq.7–9
- variational M-step: update $q(v_{dl})$ using Eq.10 and update $q(\lambda)$ using Eq.11

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

- Rectified

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

- binary

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

- nonnegative integer

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

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

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

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

The variational posterior of λ_m is,

$$q(\lambda) = \mathcal{G}\left(N/2\eta_{dl}, \left(h_{dl} + \tau\right)/2\right). \tag{11}$$

大規模メタアナリシス (学習済み word2vec のように配布), 因果推論（g-formula）, 時空間相関