大規模な疎行列に適した確率的主成分分析の計算法とその拡張

阿部興 [1] (発表者) ・島村徹平 [1]

2026 年 6 月 13 日

[1] 東京科学大学 総合研究院 難治疾患研究所

動機: 事例

細胞ごとに遺伝子の発現量を計測できる 単一細胞 RNA-seq は疾患の機序を解明するための豊富な情報を持つ.

しかし高次元で 0 が多い（疎；sparse）

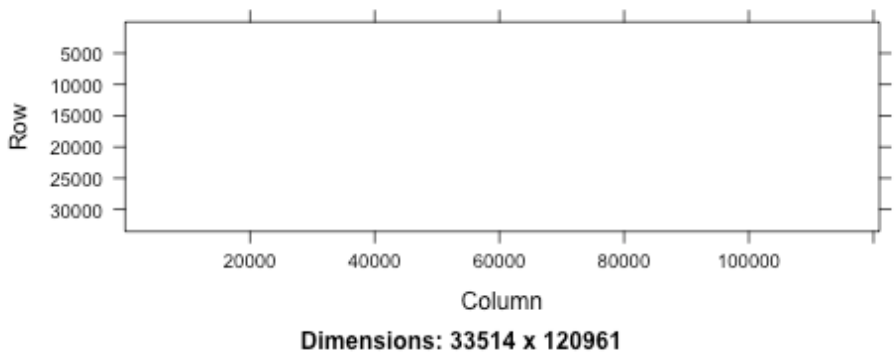

Fig: ゼロ要素を白, 非ゼロ要素を黒としたヒートマップ. 黒はほとんど見えない. 各行は遺伝子，各列は細胞. Bischoff et al. (2021)[1].

[1] Bischoff, P., Trinks, A., Obermayer, B. et al. Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. Oncogene 40, 6748 – 6758 (2021). https://doi.org/10.1038/s41388-021-02054-3

動機: 精神面

Curse of dimensionality $\longrightarrow$ Blessing of dimensionality[1]

- ◯ 高次元の積分は難しい
- △ 高次元のデータはややこしい
- ◎ 見えるものが多いとうれしい
  たとえそのほとんどが “0” だとしても

大規模で疎なデータを活用するためには…

- 計算効率を高める
- 背景知識の活用

[1] Gelman, A. (2004). “The blessing of dimensionality”
https://statmodeling.stat.columbia.edu/2004/10/27/the_blessing_of/

主成分分析

$$\underbrace{Y}_{D_1 \times D_2} \approx \underbrace{V^{(1)}}_{D_1 \times L} \underbrace{\{V^{(2)}\}^\top}_{L \times D_2}$$

他の分析の入力としても使われる

疎行列の形式

Coordinate (COO) 形式：
行列 Y の (i,j) 成分 y_{ij} を
$\mathcal{D}_n = (\boldsymbol{x}_n, y_n) = ((x_{n1}, x_{n2}), y_n) = ((i,j), y_{ij})$ と表す.
$y_{ij} = 0$ となる (i,j) は省略し，$\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_{N_1})$ とする.

Tab: COO 形式：B は A と同じ情報を持つ

Tab: A

1	0	2
0	0	2
4	1	0

Tab: B

x_{n1}	x_{n2}	y_n
1	1	1
3	1	4
3	2	1
1	3	2
2	3	2

十分統計量再考

指数型分布族：

$$p(y|\theta) = A(y)\exp(\underbrace{T(y)'}_{\text{十分統計量}} \cdot \underbrace{\eta(\theta)}_{\text{自然パラメータ}} - \underbrace{\beta(\theta)}_{\text{対数分配関数}} \cdot \gamma).$$

十分統計量 $T(0) = 0$ のとき,

$$\sum_i \log p(y_i|\theta)$$
$$= \left(\sum_{i \in \text{nonzero part}} T(y_i)'\eta(\theta) \right) - \left(\sum_{j \in \text{all of the data}} \beta(\theta)\gamma \right).$$

ゼロ要素にアクセスせずに対数尤度の評価ができる.

ベイズ更新

指数型分布族：

$$p(y|\theta) = A(y)\exp(\underbrace{T(y)^\top}_{\text{十分統計量}} \cdot \underbrace{\eta(\theta)}_{\text{自然パラメータ}} - \underbrace{\beta(\theta)}_{\text{対数分配関数}} \cdot \gamma)$$

共役事前分布：

$$\phi(\theta|\xi_1,\xi_0) = \frac{1}{Z(\xi_1,\xi_0)}\exp(\underbrace{\eta(\theta)^\top}_{\text{自然パラメータ}} \cdot \xi_1 - \underbrace{\beta(\theta)}_{\text{対数分配関数}} \cdot \xi_0)$$

事後分布：

$$\begin{aligned}\phi^*(\theta|\xi_1,\xi_0,y) \ (&\propto \phi(\theta|\xi_1,\xi_0)p(x|\theta)) \\ &= \phi(\theta|\xi_1 + \underbrace{T(y),}_{\text{十分統計量}} \quad \xi_0 + \gamma).\end{aligned}$$

$T(0) = 0$ ならば，ゼロ要素にアクセスせず incremental に計算できる．

例: 正規分布

$$\mathcal{N}(y|\mu,\sigma^2) = A(y)\exp\left((y,y^2)\begin{pmatrix}\mu/\sigma^2\\(2\sigma^2)^{-1}\end{pmatrix} - (-2\mu^2/\sigma^2)\right).$$

事前分布を正規分布 $\mathcal{N}(\mu|0,\tau^{-1})$ とすると，分散 σ^2 を所与としたときの μ の事後分布は,

$$\phi^*(\mu|y) = \mathcal{N}(t/(h+\tau\sigma^{-2}),(h+\tau)^{-1}) \text{ where}$$
$$t = y\sigma^{-2}, \quad (\mu \text{ の項の係数})$$
$$h = \sigma^{-2} \quad (\mu^2 \text{ の項 の係数})$$

疎行列の行列分解に適した変分ベイズ法

変分事後分布：独立性を仮定

$$q(\theta) = \underbrace{q(V^{(1)})}_{\text{行}}\underbrace{q(V^{(2)})}_{\text{列}}\underbrace{q(U)}_{\text{その他}} \quad \text{for } \theta = (U, V^{(1)}, V^{(2)}).$$

更新式（主結果）

$$q_t(V^{(i)}) \propto \phi\left(\hat{\xi}_1^{(i)}, \hat{\xi}_0^{(i)}\right) \quad \text{where}$$

$$\hat{\xi}_1^{(i)} = \sum_n T(y_n)' E_{q_t(\theta \setminus V^{(i)})}\left[\eta(\theta; x_n)\right] + \xi_1 \quad \text{and}$$

$$\hat{\xi}_0^{(i)} = \left(\sum_j E_{q_t(\theta \setminus V^{(i)})}[\gamma_k(\theta_{-k}; x_j)]\right) + \xi_0.$$

変分 EM アルゴリズム：
変分事後分布 $q(V^{(1)})$, $q(V^{(2)})$, $q(\lambda)$ の更新を順に繰り返す

モデル：主成分分析

COO 形式の $X = (x_1, \ldots, x_{N_1})^\top$ に合わせ，内積を次のように表記する：

$$f_n = \sum_{l=1}^{L} f_{nl} = \sum_{l=1}^{L} \prod_{k=1}^{2} V^{(k)}[X[n,k], l]$$

モデル：

$$y_n \sim \mathcal{N}(f_n, \lambda^{-1})$$

事前分布：

$$V[i,l] \sim \mathcal{N}(0, \tau^{-1})$$
$$\lambda \sim \mathcal{G}(a,b) \quad \text{（ガンマ分布）}$$

伝統的な主成分分析との関係

列パラメータ $V^{(2)}$ について周辺化：

$$Y[i,:] \sim \mathcal{N}\left(O, \{V^{(1)}\}^\top V^{(1)} + \lambda^{-1} I\right)$$

$$\mathcal{L}(V^{(1)}) = -\frac{D_1}{2} \operatorname{trace}\left(\left(\{V^{(1)}\}^\top V^{(1)} + \lambda^{-1}\right)\left(\frac{1}{D_1} Y'Y\right)\right)$$

$$(Y'Y)V^{(1)} = \overbrace{\tau}^{\text{固有値}} \quad \underbrace{V^{(1)}}_{\text{固有ベクトル}}$$

非負値行列因子分解との関係

事前分布を切断正規分布に変更すると非負制約の下での行列分解モデルが得られる.

$$V^{(k)}[i,l] \sim \mathcal{TN}(0,\tau), \quad \text{（0 以上の範囲に切断された正規分布）.}$$

note: 非負制約の有無は変数ごとに選択可能.

モデルの対数尤度

$$\mathcal{L}(V^{(k)}[d,l]) = -\sum_{n\in\delta_1}\frac{\lambda}{2}(y_n - \underbrace{f_n}_{\text{内積}})^2 + C$$

$$= -\frac{h_{kl}}{2}\left((V^{(k)}[d,l])^2 - 2V^{(k)}[d,l]\frac{t^{(k)}_{dl}}{h_{kl}}\right) + C.$$

C は $V^{(k)}$ に依存しない定数項,

$$\text{(1 次の項)}\quad t^{(k)}_{dl} = \sum_{n\in\{n|X[n,k]=d\}} V^{(k)}[X[n,k],l]\left(y_n - \sum_{l'\neq l} f_{nl}\right),$$

$$\text{(2 次の項)}\quad h_{kl} = \lambda\sum_{d=1}^{D_k}(V^{(k')}[d,l])^2, \quad \text{where } k'\neq k.$$

変分事後分布：$q(V^{(k)}[d,l])$

変分事後分布による v の期待値を $\langle v \rangle$ と書く.

$$
\begin{aligned}
&q(V^{(k)}[d,l]) \\
&= \begin{cases} \mathcal{N}(V^{(k)}[d,l] \mid \hat{\mu}_{dl}^{(k)}, \hat{\sigma}_{kl}^2) & \text{if the prior of } V^{(k)}[d,l] \text{ is not truncated} \\ \mathcal{TN}(V^{(k)}[d,l] \mid \hat{\mu}_{dl}^{(k)}, \hat{\sigma}_{kl}^2) & \text{if the prior of } V^{(k)}[d,l] \text{ is truncated,} \end{cases}
\end{aligned}
$$

変分事後分布のパラメータ $\hat{\mu}_{dl}$, $\hat{\sigma}_{dl}$ は,

$$
\begin{aligned}
\hat{\mu}_{dl}^{(k)} &= \frac{\langle t_{dl}^{(k)} \rangle}{\langle h_{kl} \rangle + \tau / \langle \lambda \rangle}, \\
\hat{\sigma}_{kl}^2 &= \left(\tau + \langle h_{kl} \rangle \right)^{-1}.
\end{aligned}
$$

変分事後分布：$q(\lambda)$

λ の変分事後分布は，形状パラメータ $\hat{a}$，レートパラメータ $\hat{b}$ のガンマ分布で与えられる．

$$q(\lambda) = \mathcal{G}(\hat{a}, \hat{b}).$$

$\hat{a}$, $\hat{b}$ は，

$$\hat{a} = N/2,$$

$$\hat{b} = \frac{1}{2}\left(\left\{\sum_{n\in\delta_1} y_n^2 + y_n\langle f_n\rangle\right\} + \left\{\sum_{i,j}\langle(V^{(1)}[i,l])^2\rangle\langle(V^{(2)}[j,l])^2\rangle\right\} + \left\{\sum_{i\neq j}\langle f_{il}\rangle\langle f_{jl}\rangle\right\}\right).$$

N：ゼロ要素も含めた全要素の数

擬似コード: V の更新

1: Function $t_{dl}^{(k)}$ の更新
2: $t_{dl}^{(k)}$ を初期化
3: **for** $n \in \{1, \ldots, N_1\}$ **do** ▷ ゼロ要素にアクセスしていない
4: $\quad t_{dl}^{(k)} \leftarrow t_{dl} + \langle V^{(k)}[d, l] \rangle (y_n - \langle f_n - f_{nl} \rangle)$
5: **end for**
6: EndFunction

1: Function h_{kl} の更新
2: h_{kl} を初期化
3: $h_{kl} \leftarrow b + \sum_j \langle (V^{k'}[j, l])^2 \rangle \ (k' \neq k)$ ▷ $\mathcal{D}$ にアクセスしていない
4: EndFunction

データ分析: Bischoff et al. (2021)

肺がんに関する単一細胞 RNA 発現量データ（再掲）

- 行（遺伝子）: 33, 514
- 列（細胞）: 120,961
- 非ゼロ要素: 239,634,370 (全体の 5%程度)

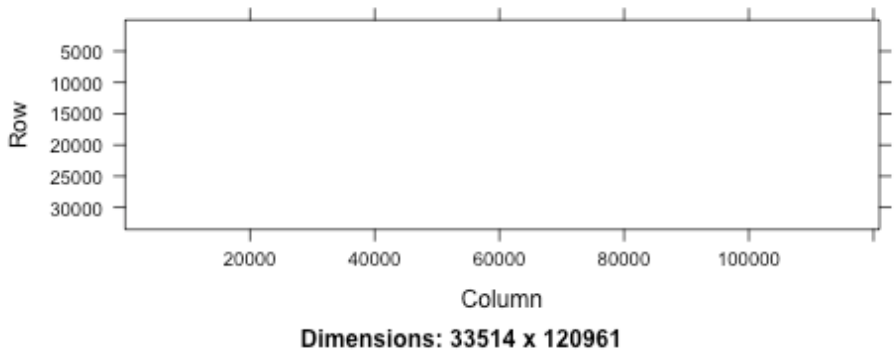

Fig: ゼロ要素を白, 非ゼロ要素を黒としたヒートマップ. 黒はほとんど見えない.

すべてをメモリ上に展開するのは無理がある

結果

細胞（列）の特徴量 W をプロット

拡張: Blessing of dimensionality へ向けて

背景知識を用いてグループ化が可能なとき

Tab: 行・列に注釈のついた行列. A 表は B 表のように表せる.

Tab: A

a	b	b	
1	0	2	A
0	0	2	A
4	1	0	B

Tab: B

x_{n1}	x_{n2}	x_{n3}	x_{n4}	y_n
1	1	a	A	1
3	1	a	B	4
3	2	b	A	1
1	3	b	A	2
2	3	b	B	2

モデルは $K > 2$ へ自然に拡張：

$$f_n = \sum_{l=1}^{L} f_{nl} = \sum_{l=1}^{L} \prod_{k=1}^{K} V^{(k)}[X[n,k],l]$$

まとめと議論

疎行列の行列分解に適した変分ベイズ法を提案した.
大規模で疎なデータを活用するためには…

- 計算効率を高める
- 背景知識の活用

本研究はそのための具体的方法の確立を目指すもの.

発展：

- 時間・空間的データ
  - 正規分布は時間・空間的に自己相関を持つ事前分布が積極的に研究されている

欠損値のある場合

$$q_t(V^{(i)}) \propto \phi\left(\hat{\xi}_1^{(i)}, \hat{\xi}_0^{(i)}\right) \quad \text{where}$$

$$\hat{\xi}_1^{(i)} = \sum_n T(y_n)' E_{q_t(\theta \setminus V^{(i)})}\left[\eta(\theta; x_n)\right] + \xi_1 \quad \text{and}$$

$$\hat{\xi}_0^{(i)} = \left(\sum_j E_{q_t(\theta \setminus V^{(i)})}[\gamma_k(\theta_{-k}; x_j)]\right) + \xi_0.$$

w_j は経験分布を用いる. 独立性は仮定.