# 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

東京科学大学総合研究院難治疾患研究所　阿部興
東京科学大学総合研究院難治疾患研究所　島村徹平

## 1　はじめに

主成分分析は多変量解析の手法として知られ，持つ情報をなるべく維持しながらデータを低次元に圧縮する目的でよく用いられる．近年は UMAP (Van der Maaten and Hinton, 2008) や t-SNE (Diaz-Papkovich et al., 2019) といった非線形の次元削減手法も広く知られるが，これらの利用にあたっても，主成分分析によって得られた主成分を入力とすることが多く，主成分分析の重要性は未だ失われていない．結果の解釈が比較的容易な点も主成分分析の利点である (Chari and Pachter, 2023).

生命科学分野においては特に，次元削減の対象としたい高次元のデータはゼロ要素が多い行列であることが多い．一細胞 RNA-seq の結果として得られる細胞 × 遺伝子発現量の行列はその代表的な例である (e.g. Lobato-Moreno et al., 2025) . 要素にゼロが多い行列は疎行列と呼ばれ，高次元の疎行列に対してはメモリ効率の観点から適した形式がある．一方で，確率モデルや推定論の観点からは，データが疎であることを積極的に活用して計算の効率を高めようとする議論はあまりされてこなかった．そこで本報告では，行列因子分解について，ゼロ要素へのアクセスを省略して計算効率を高める方法を考察する．

## 2　手法

### 2.1　疎行列と尤度関数

まず, 本報告で考察の対象とする疎行列のデータ構造について述べる．疎行列にはいくつかの異なるデータ形式がある．ここでは, 行列 $Y = (y_{ij})$ の $(i, j)$ 要素を, 変数の組 $\mathcal{D}_n = ((i, j), y_{ij}) = ((x_{n1}, x_{n2}), y_n) = (\boldsymbol{x}_n, y_n)$ として表現する．$y_{ij} = 0$ の場合, $(i, j)$ は省略し, 非ゼロ要素のみを記録するものとする．$\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_{N_1})$ とし，$N_1$ を行列 $Y$ の非ゼロ要素の数とする．この例を表 1 に示す．疎行列の表現は一様ではなく，他にも複数の方法がある．この形式は Coordinate (COO) 形式と呼ばれる．COO 形式は `R` の `Matrix` パッケージのクラス `TsparseMatrix` として実装されている形式と同等であり，テキストファイルで疎行列を保存するためによく使用される Matrix Market 形式 (Bates et al., 2024) でも採用されている実用的なものである．

次に，指数分布族の尤度について考える．ベイズ法または最尤法を用いて未知パラメータ $\theta$ を推定する場合，尤度を評価する必要がある．そしてどのようなモデルであっても，十分統計量を知ることで尤度関数を決定することができる．そこで，十分統計量の概念を改めて考える．

指数分布族は次式のように表される確率分布として定義される．

$$p(y; \theta) = \psi(y) \exp(T(y)' \eta(\theta) - \beta(\theta))$$

ここで，ベクトル値関数 $T(y)$ と $\eta(\theta)$ はそれぞれ十分統計量，自然パラメータと呼ばれる．$y$ に依存しない関数 $\beta(\theta)$ は対数正規化関数，または対数分配関数と呼ばれる．

COO 形式の $\boldsymbol{x}_n$ が与えられた下で，$y_n$ $(n = 1, \ldots, N_1)$ は互いに条件付き独立であると仮定する．さらに $T(0) = (0, \ldots, 0)'$ とすると，指数分布族の行列分解モデルについて，対数尤度 $\mathcal{L}(\theta)$ を次のように書くことが

表 1: 疎行列の例．左の表と右の表は同じ情報を持つ．右表の COO 形式を本報告で考察の対象とする．

| | | |
|---|---|---|
| 1 | 0 | 2 |
| 0 | 0 | 2 |
| 4 | 1 | 0 |

| $x_{n1}$ | $x_{n2}$ | $y_n$ |
|---|---|---|
| 1 | 1 | 1 |
| 3 | 1 | 4 |
| 3 | 2 | 1 |
| 1 | 3 | 2 |
| 2 | 3 | 2 |

できる．

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{i=1}^{N} \log p(y_i;\theta) \\ &= \left(\sum_{i\in\delta_1} T(y_i)'\eta(\theta;\boldsymbol{x}_i))\right) + \left(\sum_{\boldsymbol{x}} \beta(\theta;\boldsymbol{x})\right),\end{aligned}$$

ここで，$\delta_1$ はゼロ要素を省略した COO 形式 $\mathcal{D}$ で表されるデータセットのインデックスとした．第 1 項はデータの非ゼロ部分にのみ依存し，第 2 項は $\boldsymbol{x}$ の取りうる値の範囲のみに依存し， $y_n$ に依存しない．従い，データのゼロ部分を避けて対数尤度 $\mathcal{L}(\theta)$ を評価できる．

指数型分布族は共役事前分布を持つ．一般には，行列分解のような指数型分布族の混合は指数型分布族とは限らない．しかし，条件付き共役事前分布を使用すると，平均場近似を通じて変分ベイズ法 (Blei et al., 2017) を問題に適用できる．行または列についてのパラメータ $\theta_k$ の，$\theta_k$ を除くすべての変数で条件付けたとき共役になる事前分布は，ハイパーパラメータ $\xi = (\xi_1, \xi_0)$ を用いて次式のように表すことができる．

$$\phi(\theta_k|\xi) = \frac{1}{Z(\xi)} \exp(\eta(\theta_k)\xi_1 + \xi_0).$$

ここで $1/Z(\xi)$ は正規化のための定数である．条件付き共役事前分布 $\phi(\theta_k|\xi)$ を用いると，条件付きの事後分布は次式のように表すことができる．

$$\begin{aligned}\phi^*(\theta_k|\xi, y, \{\theta \setminus \theta_k\}) &= \frac{\phi(\theta_k|y, \{\theta \setminus \theta_k\})p(y|\theta)}{\int \phi(\theta_k|y, \{\theta \setminus \theta_k\})p(y|\theta)d\theta} \\ &= \phi(\theta_k|\xi_1 + T(y), \xi_0 + \beta_{k'}(\theta_{k'})) \quad \text{where } k' \neq k\ .\end{aligned}$$

右辺条件部の $\beta_{k'}(\theta_{k'})$ は $y$ の値に依存せず，$T(y)$ は非ゼロ要素のみに依存する．この事実は，疎行列のゼロ要素にアクセスすることなく変分ベイズ法を実行できることを意味する．

## 2.2 モデルと推定方法

これより具体的な問題として主成分分析を考える．Tipping and Bishop (1999) に倣い，主成分分析を確率モデルとして定式化することから始める．以降，添字が多重に付くことが増えるので，見やすさのために行列 $V$ の $(i,j)$ 成分を $V[i,j]$ と書くことにする．また，$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_1})'$ と置く．COO 形式にあわせ，内積を次のように表記することにする．

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

主成分分析は次のモデルを考えることと等価である．

$$y_n \sim \mathcal{N}\left(f_n, \lambda^{-1}\right), \tag{2.1}$$

ここで $\mathcal{N}(\mu,\sigma^2)$ は平均 $\mu$，分散 $\sigma^2$ の正規分布を表す．本研究では次のように，未知パラメータ $V^{(k)}$ の要素それぞれに独立な正規事前分布を置く．

$$V^{(k)}[i,l] \sim \mathcal{N}(0,\tau), \quad (i = 1,\ldots,D_k; k = 1,2).$$

あるいは，次のように事前分布を切断正規分布に変更することで非負制約の下での行列分解モデルが得られる．

$$V^{(k)}[i,l] \sim \mathcal{TN}(0,\tau), \quad (i = 1,\ldots,D_k; k = 1,2).$$

ここで $\mathcal{TN}(\mu,\sigma^2)$ は平均 $\mu$，分散 $\sigma^2$ の正規分布を 0 以上の範囲に制限した切断正規分布を表す．この場合モデルは 2 乗誤差を採用した場合の非負値行列因子分解 (Lee and Seung, 2000) に相当する．いずれの場合

も，精度パラメータ $\lambda$ に対しては次の形状パラメータ $a$，レートパラメータ $b$ のガンマ事前分布を置き，これを $\lambda \sim \mathcal{G}(a, b)$ と表記する．

これより，変分ベイズ法に基づいて潜在変数の推定量を導出する．モデル (2.2) の対数尤度を $V^{(k)}[d, l]$ に関する部分にのみ着目すると次のように表すことができる．

$$\begin{aligned}\mathcal{L}(V^{(k)}[d,l]) &= \sum_{n\in\delta_1} \log p(y_n|V, X, \lambda) \\ &= \sum_{n\in\delta_1} \left(-\frac{\lambda}{2}(y_n - f_n)^2\right) + C \\ &= -\frac{h_{kl}}{2}\left((V^{(k)}[d,l])^2 - 2V^{(k)}[d,l]\frac{t^{(k)}_{dl}}{h_{kl}}\right) + C.\end{aligned}$$

ここで $C$ は $V^{(k)}$ に依存しない定数項であり，$t^{(k)}_{dl}$ と $h_{kl}$ は次のように定める．

$$\begin{aligned}t^{(k)}_{dl} &= \sum_{n\in\{n|X[n,k]=d\}} V^{(k)}[X[n,k],l]\left(y_n - \sum_{l'\neq l} f_{nl}\right), \\ h_{kl} &= \lambda\sum_{d=1}^{D_k}(V^{(k')}[d,l])^2, \quad \text{where } k' \neq k.\end{aligned}$$

2.1 節で考察したように，$t^{(k)}_{dl}$ は非ゼロ要素のみから漸進的（incremental）に計算することが可能であり，$h_{kl}$ は行列の要素の取る値に依存しない．平均場近似，すなわち変分事後分布がすべてそれぞれ互いに独立であるとした下で，変分事後分布における $v$ の期待値を $\langle v\rangle$ と書くことにすると，$V^{(k)}[d, l]$ の変分事後分布は次式のようになる．

$$q(v_{dl}) = \begin{cases}\mathcal{N}(v_{dl} \mid \hat{\mu}_{dl}, \hat{\sigma}_{dl}) & \text{if the prior of } v_{dl} \text{ is not truncated} \\ \mathcal{TN}(v_{dl} \mid \hat{\mu}_{dl}, \hat{\sigma}_{dl}) & \text{if the prior of } v_{dl} \text{ is truncated,}\end{cases} \tag{2.2}$$

ここでは $\hat{\mu}_{dl}$ と $\hat{\sigma}_{dl}$ をそれぞれ次のように定める．

$$\hat{\mu}_{dl} = \frac{\langle t_{dl}\rangle}{\langle h_{dl}\rangle + \tau/\langle\lambda\rangle}, \quad \hat{\sigma}^2 = (\tau + \langle h_{dl}\rangle)^{-1}.$$

$\lambda$ 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．

$$q(\lambda) = \mathcal{G}(\hat{a}, \hat{b}). \tag{2.3}$$

$\hat{a}, \hat{b}$ は，$N$ をゼロ要素も含めた全要素の数として，それぞれ次のように定める．

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

変分ベイズ法のアルゴリズムは $V^{(k)}$ $(k = 1, 2)$ の変分事後分布 (2.2) と $\lambda$ の変分事後分布 (2.2) を順に更新する反復によって得られる．更新式に必要な変分事後分布による期待値は次のリストに示す．

- $\langle v_{dl}\rangle = \begin{cases}\hat{\mu}_{dl} + \hat{\sigma}_{dl}\phi(-\hat{\mu}_{dl}/\hat{\sigma}_{dl})/\Phi(-\hat{\mu}_{dl}/\hat{\sigma}_{dl}) & \text{if the prior of } v_{dl} \text{ is truncated,} \\ \hat{\mu}_{dl} & \text{if the prior of } v_{dl} \text{ is not truncated}\end{cases}$
- $\langle v^2_{dl}\rangle = \begin{cases}\hat{\mu}^2_{dl} + \hat{\sigma}^2_{dl} + \hat{\mu}_{dl}\hat{\sigma}_{dl}\phi(-\hat{\mu}_{dl}/\hat{\sigma}_{dl})/\Phi(-\hat{\mu}_{dl}/\hat{\sigma}_{dl}) & \text{if the prior of } v_{dl} \text{ is truncated} \\ \hat{\mu}^2_{dl} + \hat{\sigma}^2_{dl} & \text{if the prior of } v_{dl} \text{ is not truncated}\end{cases}$
- $\langle\lambda\rangle = \hat{a}/\hat{b}$.

表 2: 行・列に注釈のついた行列. 左の表は右の表のように COO 形式の拡張で表せる.

| a | b | b | |
|---|---|---|---|
| 1 | 0 | 2 | A |
| 0 | 0 | 2 | A |
| 4 | 1 | 0 | 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 |

ここで $\phi(x)$ と $\Phi(x)$ はそれぞれ標準正規分布の密度関数と分布関数である．結果からわかるように，非負制約の有無は対応する変数ごとに設定することが可能である．このことは変数によって分布の台が明確に変わるデータ——ある変数は正の値のみを取り，ある変数は負の値を取り得るような場合——を分析する際に役立つ.

紙幅の都合により，本手法の計算効率についてのシミュレーションとデータ分析事例は当日の発表で示す.

## 3 まとめと議論

本報告では，スパースデータに適した行列分解のアルゴリズムを考察した．データのスパース性は計算コストとメモリコストの削減につながる．大規模データの時代においては，十分統計量が計算コスト削減の観点から見直される可能性がある.

最後に，本研究の拡張について述べておく．すぐにわかるように，COO 形式は 2 次元以上の多次元配列へ容易に拡張できる（表 2）．応じて 式 (2.2) のモデルは $K > 2$ に拡張できる．この場合，モデルは CANDECOMP/PARAFAC (CP) 分解と呼ばれるテンソル分解の手法と同等になる．加えて，背景知識に基づいて表 2 のように変数をグループ化できる場合，変数の数が増えるほどグループごとのパラメータを推定するためのサンプルは増える．この場合，高次元であることは弱点とならない.

## 参考文献


D. Bates, M. Maechler, and M. Jagan. Matrix: Sparse and dense matrix classes and methods, 2024. URL `https://CRAN.R-project.org/package=Matrix`.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19 (8):e1011288, 2023.

Alex Diaz-Papkovich, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, 15(11): e1008432, 2019.

Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

Sergio Lobato-Moreno, Umut Yildiz, Annique Claringbould, et al. Single-cell ultra-high-throughput multiplexed chromatin and rna profiling reveals gene regulatory dynamics. *Nature Methods*, 2025. doi: 10.1038/s41592-025-02700-8. URL `https://doi.org/10.1038/s41592-025-02700-8`.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579–2605, 2008.