分離可能想定下の非負行列分解に対する組合せ的解法の開発と応用

静岡大学工学部数理システム工学科
水谷　友彦

１．はじめに

分離可能想定下の非負行列分解（Separable NMF）は画像処理、テキストマイニング、クラウドソーシングなどに応用を持つ。本研究課題ではSeparable NMFに対して組合せ的最適化問題に基づく手法とその緩和に基づく手法について研究を行った。

２．分離可能想定下の非負行列分解

非負行列$A \in R_+^{d\times n}$は二つの非負行列$W \in R_+^{d\times r}, H \in R_+^{r\times n}$を用いて、$A = WH$というように積の形に分解できるとする。このような分解は**非負行列分解**（NMF）と呼ばれる。Aの NMF $A = WH$において行列Hが次のような条件を満たすとき、

$$H = [I, \bar{H}]\Pi \in R^{r\times n}$$

Aの NMF は**分離可能**(separable)であると呼ばれる。ここで、Iはr次単位行列、Πはn次置換行列、$\bar{H}$は大きさが$r \times (n-r)$の非負行列である。今後の議論のために用語を準備する。非負行列$A \in R_+^{d\times n}$が以下のように分解できるとき、

$$A = WH \in R_+^{d\times n},\ H = [I, \bar{\mathrm{H}}]\Pi \in R_+^{r\times n}$$

Aは**r−分離可能**であると呼ぶことにする。また、rを**分解ランク**、Wを**基底行列**と呼ぶことにする。Aの NMF がr−分離可能の場合、基底行列Wの各列はAの列の中に含まれることを意味する。したがって、Aの列の中から、基底行列の列を全て見つけることができれば、Aの NMF を計算することが可能である。

Separable NMF 問題を以下のように定める。

問題
r−分離可能な$A \in R^{d\times n}$と分解ランクrが与えられたとき、Aの基底行列を求めよ。

一般に NMF を計算することは NP 困難であるが、Separable NMF 問題は多項式時間で解くことができる。NMF に分離可能性を仮定すると、応用の範囲は限定されてしまうが、画像処理、テキストマイニング、クラウドソーシングなどに応用がある。このような応用から生じる Separable NMF はノイズを含むことを想定するのが自然である。 そこで、r−分離可能な$A \in R_+^{d\times n}$はノイズ$V \in R^{d\times n}$を含むという設定を考える。

$$A = WH + V = W[I, \bar{H}]\Pi + V$$

このようなノイズを含むようなAに対して、Separable NMF 問題に対するアルゴリズムを適用したとき、基底行列Wをよく近似する出力が得られる場合、アルゴリズムは**ノイズに対して頑強**であると言う。Separable NMF の実問題への応用においては、アルゴリズムのノイズ頑強性が成功の鍵となる。

３．直接法:組合せ的最適化問題に基づく手法

Separable NMF 問題は以下のように組合せ的最適化問題として定式化できる。

最小化　$||A - AX||$　条件　$X \in R^{n \times n}$は非負行列でr本の非ゼロな行を持つ　(1)

行列$X \in R^{n \times n}$がこの問題の変数である。問題(1)に対して、Gillis は以下のような最適化問題を解くことを提案した[5]。

最小化　$f(\mathcal{K})$　条件　$|\mathcal{K}| = r$　(2)

集合 $\mathcal{S} = \{\boldsymbol{h} \in R^r \mid \mathbf{1}^T\boldsymbol{h} \leq 1, \boldsymbol{h} \geq 0\}$ に対して関数fを以下のように定める。

$$f(\mathcal{K}) = \max_{j=1,\ldots,n} \min_{\boldsymbol{h} \in \mathcal{S}} ||A(:,j) - A(:,\mathcal{K})\boldsymbol{h}||_2$$

問題(2)の最適解$\mathcal{K}^*$に対して、$A(\mathcal{K}^*)$を構築するとAの基底行列をよく近似することを Gillis は証明した[5]。以下に結果の概要を述べる。もし、ノイズVはどんなjに対しても $||V(:,j)||_2$の値が小さくなるならば、$A(\mathcal{K}^*)$の列を適切に並び替えると基底行列Wとのギャップ $||W - A(\mathcal{K}^*)||_2$はある定数で上から抑えることができる。更に、この上限を本質的に改善することはできない。したがって、この手法は Separable NMF 問題に対しての最適な手法である。

Gillis の研究では具体的なアルゴリズムを示してない。そこで、本研究では 0/1 整数計画問題に基づくアルゴリズムを開発し、Gillis の研究と同様の理論結果を証明することに成功した。ここでは、0/1 整数計画問題に基づく提案手法について説明する。問題(1)において目的関数のノルムを 1 ノルムに定めると、この問題は以下のように書き換えることができる。

最小化　$||A - A(\mathcal{K})Y||_1$　条件　$Y \geq O, |\mathcal{K}| = r$　(3)

行列$Y \in \mathrm{R}^{r \times n}$と添字集合$\mathcal{K}$が問題の変数である。この問題は以下のような 0/1 整数計画問題に帰着することができる。

$$
\begin{aligned}
\text{(P)}\quad &\text{最小化} && z \\
&\text{条件} && A - AX = F - G \\
& && \textstyle\sum_{i=1}^{n} F(i,j) + G(i,j) \le z,\ j = 1,\dots,n \\
& && \textstyle\sum_{j=1}^{n} X(i,j) \le M s_i,\ i = 1,\dots,n \\
& && s_1 + \cdots + s_n = r \\
& && X \ge O, F \ge O, G \ge O \\
& && s_i \in \{0,1\}, \quad i = 1,\dots,n
\end{aligned}
$$

変数は$(X,F,G,s,z) \in R^{n\times n} \times R^{d\times n} \times R^{d\times n} \times \{0,1\}^n \times R$である。$M$は十分大きな実数とする。今、$A$は$r$−分離可能であるとしよう。このとき問題(P)の最適解(X^*,F^*,G^*,s^*,z^*)から添字集合$\mathcal{K} = \{\, i \mid s_i^* = 1 \,\}$を構築すると、$A$の基底行列は$A(\mathcal{K})$と一致する。これより以下のようなアルゴリズムを設計できる。

アルゴリズム 1
入力：$A \in R^{d\times n}$と正数 r
出力：要素数 rの添字集合$\mathcal{K}$

1: (A,r)に関して問題(P)を解いて最適解(X^*,F^*,G^*,s^*,z^*)を求める。
2: 添字集合$\mathcal{K} = \{\, i \mid s_i^* = 1 \,\}$を構築し出力する。

アルゴリズム 1 を MATLAB で実装した。0/1 整数計画問題のソルバーとして CPLEX を利用した。人工的にデータを生成し、ノイズ頑強性と計算時間を調べた。その結果、提案手法はノイズ頑強性に優れていることを確認した。また、$n \le 200$ 程度ならば現実的な時間でアルゴリズムは終了することを確認した。予想通り、ノイズ頑強性の点では優れているが、計算時間に課題があることが分かった。

4. 緩和法：組合せ的最適化問題の緩和に基づく手法

4−1. アルゴリズム

組合せ的最適化問題の緩和に基づく手法について考えよう[7]。以下のように問題(3)の緩和問題(H)を構築する。

(H)　最小化　$|| A - AX ||_1$

条件　$\mathrm{tr}(X) = r$

$0 \le \mathrm{X(i,j)} \le \mathrm{X(i,i)} \le 1, \quad \mathrm{i,j} = 1, \dots, \mathrm{n}$

行列$X \in R^{n \times n}$がこの問題の変数である。問題(H)は線形計画問題に帰着することができる。アルゴリズム 1 において(P)の代わりに(H)を解くアルゴリズムを設計する。

アルゴリズム 2

入力：$A \in \mathrm{R}^{\mathrm{d} \times \mathrm{n}}$と正数 r

出力：要素数rの添字集合$\mathcal{K}$

1: (A, r)に関して問題(H)を解いて最適解X^*を求める。

2: X^*の対角要素を大きい順からr個選び、その添字集合$\mathcal{K}$を出力する。

アルゴリズム 2 はノイズに対して頑強であることが理論的に示されている。しかし、問題(H)は問題(P)に比べて解くことが容易であるが、行列の規模が大きくなると(H)を解くためには多くの計算時間が必要となる。そのため、(H)に対して効率的解法を開発する必要がある。本研究では列生成法に基づく手法を開発した。その概要を以下で説明する。

(H)の実行可能解Xは多くの零要素を持つ。実際、制約式 $0 \le X(i,j) \le X(i,i) \le 1$ より、もし$X(i,i) = 0$ならばXのi行目の全ての要素は零となる。更に制約式 $0 \le X(i,i) \le 1,\ \mathrm{tr}(X) = r$から$X$の多くの対角要素は零となることが期待できる。したがって、Xの行を適切に並べ替えると、以下のようにXは非零行列$\bar{X} \in R^{u \times n}$と零行列に分割することができる。

$$X = \begin{bmatrix} \bar{X} \\ O \end{bmatrix} \in R^{n \times n}$$

このことを踏まえると、より規模の小さい問題を解くことで(H)の最適解が得られることが分かる。$\mathcal{U} \subset \{1, \dots, n\},\ |\mathcal{U}| = u$ に対して、

$$\text{最小化} \quad || A - A(\mathcal{U})X ||_1 \quad \text{条件} \quad X \in \mathcal{C}(u) \qquad (4)$$

を定める。ここで$\mathcal{C}(u)$は以下のような条件を満たす行列$X \in R^{u \times n}$の集合である。

$\sum_{i=1}^{u} X(i,i) = r$

$0 \le X(i,j) \le X(i,i) \le 1, \quad i = 1, \dots, u, j = 1, \dots, n$

問題(4)の変数は行列$X \in R^{u \times n}$なので(H)よりも問題規模は小さい。また、問題(4)も線形計画問題に帰着することができる。$\mathcal{U}$を適切に選んで問題(4)の最適解X^*を計算する。すると、行列

$[X;O] \in R^{n\times n}$を構築して行を並べ替えると(H)の最適解に一致することを示せる。そして、列生成の枠組みを用いるとこのような$\mathcal{U}$を求めることができる。

4-2. 実験

アルゴリズム2をMATLABで実装し、ハイパースペクトル画像の端成分抽出問題に適用した。データはUrbanデータセットを用いた。このデータセットはテキサス州のある都市をHYDICEセンサで取得したものである。これまでの研究によって、Urbanデータセットの端成分はアスファルト、草、木、屋根1、屋根2、土であることが判明している。また、各端成分のスペクトルも判明している。実験では提案手法と6つの既存手法SPA[2]、PSPA[4]、ER[6]、VCA[1]、SNPA[3]を比較した。各手法で得られた端成分スペクトルの推定値$\hat{\boldsymbol{w}}_1,\ldots,\hat{\boldsymbol{w}}_r$がどのくらい真値$\boldsymbol{w}_1,\ldots,\boldsymbol{w}_r$に近いかを評価するためにMRSA（mean-removed spectral angle）を用いた。MRSAは0から100までの値を取る。推定値が真値に近い場合、MRSAの値は小さくなり、特に、推定値が真値に一致するとMRSAの値は0となる。表1は端成分ごとのMRSA値とその平均値をまとめたものである。この表からアルゴリズム2による予測値は6つの端成分の内、5つの端成分に対して既存手法よりも真値に近いことが分かる。

表 1　端成分スペクトルごとのMRSA値と平均値

	アルゴリズム2	SPA	PSPA	ER	VCA	SNPA
アスファルト	9.2	20.5	20.5	12.7	20.5	16.2
草	5.7	18.4	18.4	50.0	35.4	15.7
木	3.4	4.7	4.7	4.7	4.7	15.3
屋根1	7.1	12.5	12.5	12.5	12.5	13.6
屋根2	18.9	29.2	29.2	17.1	17.1	50.0
土	3.2	16.7	16.7	16.7	16.7	50.0
平均	7.9	17.0	17.0	19.0	17.8	26.8

5. まとめ

本研究ではSeparable NMF問題に対して組合せ的手法を開発した。直接法のノイズ頑強性を理論的に解析した。また、人工的に生成したデータを用いて実験を行った。その結果、ノイズに対する頑強性を確認できた。また、列数が200以下ならば現実的な時間で計算が終了することが分かった。緩和法の計算効率を改善するために問題(H)に対して列生成法に基づく手法を開発した。実際に、提案手法をハイパースペクトル画像の端成分抽出問題に適用してその有効性を確認した。

謝辞

本研究は天野工業技術研究所 2023 年（特別募集）研究助成を受けて実施しました。ここに記して謝意をしまします。

参考文献

1) J. M. P. Nascimento and J. M. B. Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 43(4):898-910, 2005.
2) N. Gillis and S. A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(4):698-714, 2014.
3) N. Gillis. Successive nonnegative projection algorithm for robust nonnegative blind source separation. SIAM Journal on Imaging Sciences, 7(2):1420--1450, 2014.
4) N. Gillis and S. A. Vavasis. Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. SIAM Journal on Optimization}, 25(1):677--698, 2015.
5) N. Gillis. Nonnegative Matrix Factorization, SIAM, 2020.
6) T. Mizutani. Ellipsoidal rounding for nonnegative matrix factorization under noisy separability. Journal of Machine Learning Research, 15:1011--1039, 2014.
7) T. Mizutani. Implementing Hottopixx methods for endmember extraction in hyperspectral images. arXiv:2404.13098, 2024.