超高次元データ解析のための量子インスパイア主成分分析の開発
Quantum-inspired principal component analysis for exponentially large dimensional data

間島慶(PY)[1)]，小出（間島）真子[2)]， 八幡憲明[1)]
[1)] 量子科学技術研究開発機構
[2)] 情報通信研究機構
E-mail: majima.kei@qst.go.jp

Abstract— Principal component analysis (PCA) is a widely used statistical tool for extracting low-dimensional structures underlying multivariate data. However, its application to high-dimensional data is limited due to its large computational time. While the conventional PCA algorithm requires polynomial time, using a quantum-inspired algorithm as a subroutine, we have implemented an algorithm that approximates it with computational time proportional to the logarithm of the input dimensionality. The computational efficiency and performance of the implemented algorithm, quantum-inspired PCA, are experimentally evaluated on synthetic and real datasets.

Keywords— **principal component analysis, quantum-inspired algorithm, high-dimensional data**

1 はじめに

主成分分析は多変量データから重要な低次元成分を抽出する統計手法である. 与えられた多変量データを線形変換し, データ内の分散を最も保つ低次元成分を抽出する. 多変量データの解析において頻繁に用いられる統計手法の一つであり, 例えば神経科学の研究では多地点から計測された神経活動のデータから有用な解釈を得るために用いられる.

しかし，主成分分析は計算時間の問題から，高次元データへの適用がしばしば困難となる. 特異値分解に基づいたアルゴリズムを用いた場合, 取り扱うデータの変量の数（次元数）に対し，2乗で計算時間が増加する. 近年ではデータの計測技術の進展・高解像度化により，次元数は数千万以上に達することがあり，汎用型CPU搭載のPCを用いた場合，数週間以上の計算時間を要する.

このような計算時間の増加に対処するため, 本報告では量子インスパイアアルゴリズムを導入する. 量子インスパイアアルゴリズムは量子コンピュータの有用性を検証する過程で提案された古典アルゴリズムであり，近似にはなるが，特異値分解の計算量を次元数の対数オーダーに抑えることができる [1]. この量子インスパイアアルゴリズムを用いることで線形回帰，主成分分析，正準相関分析，非負値行列分解，サポートベクトルマシンなどを高速に近似するアルゴリズムが提案されている. しかし，近似精度に関する理論的な研究がなされる一方で, それら量子インスパイアアルゴリズムは実際のデータ解析においてまだほとんど用いられておらず, その有用性は未知数である.

本報告では, 数百万～数億次元に達する高次元データ解析に向けて, 量子インスパイアアルゴリズムによる主成分分析の計算時間と性能を計算機実験によって評価する.

2 アルゴリズム

この節では本報告の実験で用いるアルゴリズムの概要を説明する. 2.1節で主成分分析, 2.2節で量子インスパイアアルゴリズムを用いた主成分分析について説明する.

2.1 主成分分析

サンプル数 N，次元数（変量の数）D のデータ行列 $\mathbf{X} \in \mathbb{R}^{N\times D}$ が与えられたとする. データのセンタリング（平均を0にシフトする前処理）はすでになされているものとする. 主成分分析では，第一主成分を抽出する重みベクトルとして, 以下の条件を満たすベクトル $\mathbf{w} \in \mathbb{R}^{D}$ を計算する.

$$\widehat{\mathbf{w}} = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}$$

ここで, $\|\mathbf{w}\|$ はベクトル $\mathbf{w}$ のL2ノルム（ユークリッドノルム），$\mathbf{X}^{\mathrm{T}}$は行列 $\mathbf{X}$ の転置，$\mathbf{w}^{\mathrm{T}}$ はベクトル $\mathbf{w}$ の転置を表す. 第二主成分以降の重みベクトルは，上と同型の最適化問題をすでに得られている重みベクトルと直交する条件のもと解くことで得られる. データ解析では通常, 上位の主成分のみに興味があることが多い. ここでは上位 K 個の主成分に興味があるとし，それに対する重みベクトルを $\widehat{\mathbf{w}}^{(1)},\cdots,\widehat{\mathbf{w}}^{(K)}$ と記す. 重みベクトルを計算するアルゴリズムは複数あ

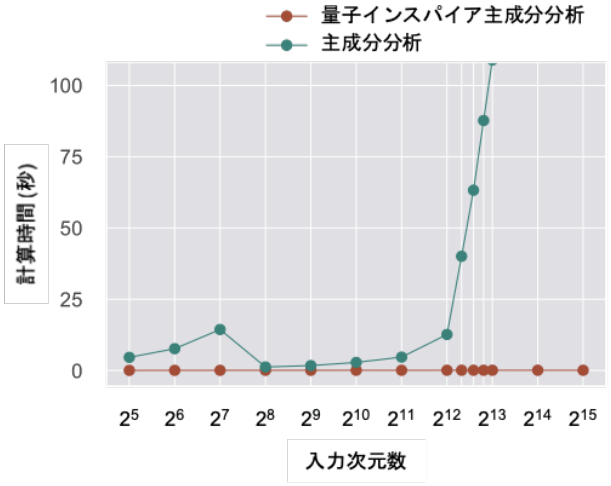

図 1. 量子インスパイア主成分分析と主成分分析の計算時間の比較. 入力次元数を変えつつ，10 回の平均計算時間をプロットした.

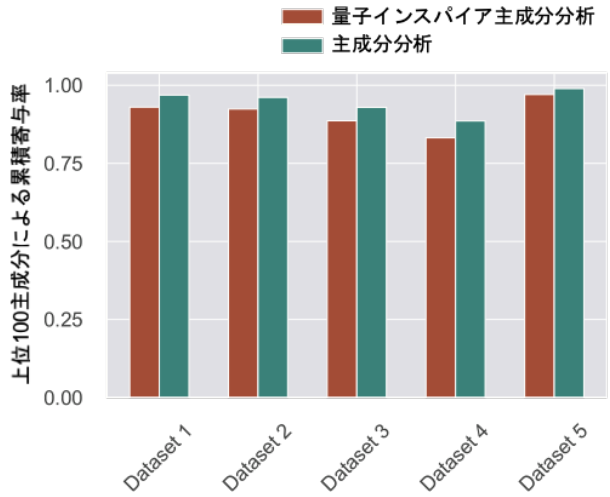

図 2. 量子インスパイア主成分分析と主成分分析の性能比較. 5 つの実データセットに適用した際の上位 100 主成分による累積寄与率を評価した.

るが，ここでは特異値分解を用いる方法を説明する. データ行列 $\mathbf{X}$ の特異値分解を $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ とおく. 第 k 番目の右特異ベクトルは第 k 主成分の重みベクトルと一致することが知られている. つまり，特異値分解の結果を用いて，$\hat{\mathbf{w}}^{(k)} = \mathbf{V}(:, k)$ として重みベクトルを求めることができる. サイズ $N \times D$ の行列の特異値分解にかかる計算量は $O(\min(N^2D, ND^2))$ であるため，サンプル数 N が十分大きい時，その計算時間は次元数 D の2乗に比例する.

2.2 量子インスパイア主成分分析

本報告の実験では先行研究 Koide-Majima & Majima [2]で使われている量子インスパイアアルゴリズムと同一のものを用いて主成分分析を行った.これは Tang [1]によって提案された特異値分解を行う量子インスパイアアルゴリズム（量子インスパイア特異値分解）をベースに作られたものである. 本報告では 2.1 節 で述べた主成分分析のアルゴリズムにおいて，特異値分解を量子インスパイア特異値分解に置き換えたものを量子インスパイア主成分分析と呼ぶ.抽出する主成分数を固定した際の計算量は $O(\log(ND))$ となるため，高次元データ解析における大幅な高速化が期待できる.

3 実験結果

最初に，シミュレーションデータを用い，主成分分析と量子インスパイア主成分分析の計算時間を比較した. ガウス乱数を用いてサイズ$N \times D$の入力行列$\mathbf{X}$を生成し，それに両アルゴリズムを適用し，計算時間を計測した. サンプル数Nは10000，入力次元数Dは$\{2^5, 2^6, \cdots, 2^{15}\}$の値をとるものとした. 図は次元数$D$の関数として計算時間をプロットしたものである. 次元数を指数関数的に増やしつつ，計算時間を評価した(図 2A).主成分分析の計算時間が指数関数的に増加していくのに対し，量子インスパイア主成分 分析では計算時間を改善することに成功している.

次に5つの実データセットを用い，量子インスパイア主成分分析の近似性能を評価した. 性能の評価指標として，抽出した上位100の主成分によって説明可能な分散の割合（累積寄与率）を用いた. 実データセットとして，MNIST，CIFAR10，WikiCLIR，JESC，XRMBを用いた（詳細は [2]を参照）. 累積寄与率による評価では，主成分分析と比べた量子インスパイア主成分分析の性能低下は最大7%程度であった.

4 考察

本報告では先行研究において提案・実装した量子インスパイア主成分分析の計算時間・性能をシミュレーションデータ・実データセットを用いて評価した. 結果，性能低下を伴うものの，数百万次元以上の高次元データを扱うのにも耐えうる程度計算時間を改善することができた. また，今回用いた実データセットにおいては，性能の低下は7%以下であった. これらの結果は量子インスパイアアルゴリズムに基づく機械学習手法によって，これまで取り扱い自体が不可能であった大規模な高次元データを解析対象にできる可能性を示唆している.

参考文献

[1] E. Tang, *A Quantum-Inspired Classical Algorithm for Recommendation Systems*, Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing - STOC 2019 217 (2019).

[2] N. Koide-Majima and K. Majima, *Quantum-Inspired Canonical Correlation Analysis for Exponentially Large Dimensional Data*, Neural Netw **135**, 55 (2021).