1C4-4

頑健性線形射影法の特性評価

Evaluation of the characteristic of Robust Linear Projection

小林 えり[*1] 伏見 卓恭[*1] 斉藤 和巳[*1] 池田 哲夫[*1]
Eri Kobayashi Takayasu Fushimi Kazumi Saito Tetsuo Ikeda

[*1]静岡県立大学
University of Shizuoka

Reducing dimensionality of high dimensional data plays an important role for understanding the intrinsic structure of data. In order to project high dimensional data to law dimensional one, the PCA (Principal Component Analysis) method has been widely used, but it may suffer from outlier samples due to its squared error criterion. In this paper, we describe and evaluate a new projection method based on an absolute error criterion, which we call the RLP (Robust Linear Projection) method. In our experiments, we compare the RLP method with the PCA and random projection methods, in terms of the values of the objective functions, the performance of similarity search, and others.

1. はじめに

近年，Ｗｅｂ上には膨大な量のデータが蓄積されており，それらを有効活用するためには，データ間の関係や特性を把握することは一層重要になっている．マルチメディアデータをはじめ, データの多くは高次元ベクトルで表現でき，高次元ベクトルで定義されたオブジェクト集合を低次元ベクトルに埋め込むことはデータの隠れた構造やオブジェクト間の関係を視覚的に把握するために重要なことである [1] [2].

さらに，蓄積された膨大な量のデータから，与えられたクエリと類似するオブジェクトを検索する類似検索の研究は，非常に重要になってきている．類似検索においては，クエリからある範囲内のオブジェクトを検索するレンジクエリ問題があるが，高次元空間から低次元空間への縮小写像を実現できれば，レンジ内に存在するオブジェクトをある程度枝刈りでき，レンジクエリ検索の効率化が図れる [3] [4].

高次元データの低次元空間への埋め込み法の代表例として，主成分分析（以下 PCA 法） [1] があげられる．PCA 法では，データの散らばり具合を最大にするような射影軸を求めるために，分散共分散行列の最大固有値を求めることで計算する．しかしデータ内には外れ値が存在することがあり，外れ値との距離が強く影響し，頑健な結果が得られない場合がある．そこで代表的な既存埋め込み法と同様に元空間でのオブジェクト間の関係をできるだけ保持した縮小写像を実現し，かつ，この外れ値に対して頑健性のある頑健線形射影法（以下 RLP 法）を提案した．RLP 法は，射影後のデータの散らばり具合として絶対値を用いて計算するため，外れ値に対して頑健であることが期待できる．

本稿では提案法 RLP 法を従来法である PCA 法に加え, 射影軸をランダムに設定したランダム射影法 (以下ランダム法) を用いて提案法の有効性を検証する．一般に射影軸をランダムに設定した射影法のほうがデータがより散らばると考えられる．2 つのレビューサイトから収集したアイテム集合をオブジェクトとして用い各手法の低次元空間への射影結果に関して，可視化結果と類似検索の観点から定量的に評価する．

連絡先: 小林 えり，静岡県立大学経営情報学部，静岡県静岡市駿河区谷田 52-1，054-264-5436，b10034@u-shizuoka-ken.ac.jp

2. 頑健線形射影法

RLP 法は，PCA 法と同様に，射影後のデータの散らばり具合が最大となる射影軸を求める手法である．以下の小節で，RLP 法のアイデアとアルゴリズム，その特徴について述べる．

2.1 提案法のアイデア

入力として元空間の座標ベクトル群 $\mathbf{Z}$ を用いる．($\mathbf{Z} = (\mathbf{z}_1 \cdots \mathbf{z}_N)$) 元空間の座標ベクトル群を射影軸 $\mathbf{w}$ に射影した値を成分とする埋め込み射影値ベクトル $\mathbf{x}$ を出力とする．射影軸 $\mathbf{w}$ は以下の目的関数 $\mathcal{R}(\mathbf{w})$ を最大化するように求める．

$$\mathcal{R}(\mathbf{w}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |\mathbf{w}^T\mathbf{z}_i - \mathbf{w}^T\mathbf{z}_j| \tag{1}$$

ここで，N は全オブジェクト数，$||\mathbf{w}|| = 1$ である．元空間ベクトル $\mathbf{Z}$ を求められた $\mathbf{w}$ に射影し，埋め込み射影値ベクトル $\mathbf{x} = \mathbf{w}^T\mathbf{Z}$ を決定する．目的関数 1 は，絶対値を用いている点で PCA 法の 目的関数 2 よりノイズに対して頑健であると期待できる．

$$\mathcal{P}(\mathbf{w}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\mathbf{w}^T\mathbf{z}_i - \mathbf{w}^T\mathbf{z}_j\right)^2 \tag{2}$$

また，目的関数 1 は，絶対値関数を最大化する問題であるため，PCA 法の場合と異なり固有値問題として解くことができない．しかし，全オブジェクトを射影後の座標値 $\mathbf{w}^T\mathbf{z}_i$ によって降順にソートすることで，目的関数を以下のように書き換えることができる．

$$\begin{aligned}
\mathcal{R}(\mathbf{w}) &= \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |\mathbf{w}^T\mathbf{z}_i - \mathbf{w}^T\mathbf{z}_j| \\
&= (N-1)\mathbf{w}^T\mathbf{z}_{r(1)} + (N-2)\mathbf{w}^T\mathbf{z}_{r(2)} - \mathbf{w}^T\mathbf{z}_{r(2)} + \cdots \\
&= (N-1)\mathbf{w}^T\mathbf{z}_{r(1)} + (N-3)\mathbf{w}^T\mathbf{z}_{r(2)} + \cdots \\
&= \mathbf{w}^T \sum_{k=1}^{N} (N-2k+1)\mathbf{z}_{r(k)}
\end{aligned}$$

表 1: 目的関数値 (L1) (@cosme)

次元	PCA	RLP	RAND	ORAND
1	5.99e+04	5.89e+04	5.61e+03	5.61e+03
2	1.07e+05	9.88e+04	1.10e+04	1.10e+04
3	1.51e+05	1.43e+05	1.66e+04	1.66e+04
10	4.08e+05	3.94e+05	5.54e+04	5.54e+04
100	2.83e+06	2.76e+06	5.48e+05	5.49e+05
1000	1.70e+07	1.68e+07	5.50e+06	5.49e+06

表 2: 目的関数値 (L2) (@cosme)

次元	PCA	RLP	RAND	ORAND
1	2.23e+04	2.28e+04	1.95e+02	1.95e+02
2	3.65e+04	3.80e+04	3.85e+02	3.85e+02
3	4.86e+04	5.12e+04	5.78e+02	5.78e+02
10	1.08e+05	1.120e+05	1.92e+03	1.92e+03
100	5.07e+05	5.13e+05	1.89e+04	1.89e+04
1000	1.96e+06	2.00e+06	1.90e+05	1.90e+05

表 3: 目的関数値 (L1) (アニコレ)

次元	PCA	RLP	RAND	ORAND
1	2.67e+05	2.726e+05	1.54e+04	1.55e+04
2	4.99e+05	5.11e+05	3.03e+04	1.10e+04
3	7.15e+05	7.27e+05	4.61e+04	4.52e+04
10	1.89e+06	1.93e+06	1.52e+05	1.50e+05
100	1.03e+07	1.05e+07	1.51e+06	1.51e+06
1000	4.59e+07	4.57e+07	1.51e+07	1.51e+07

表 4: 目的関数値 (L2) (アニコレ)

次元	PCA	RLP	RAND	ORAND
1	1.50e+05	1.46e+05	4.66e+02	4.76e+02
2	2.64e+05	2.58e+05	8.97e+02	8.75e+02
3	3.53e+05	3.47e+05	1.38e+03	1.33e+03
10	7.70e+05	7.54e+05	4.54e+03	4.44e+03
100	2.43e+06	2.39e+06	4.49e+04	4.49e+04
1000	5.54e+06	5.34e+06	4.52e+05	4.52e+05

ここで r は，射影後の値を降順ソートした順序リストであり，$r(k)$ は順位が k 番目のオブジェクトを表す．この目的関数，すなわち射影軸 $\mathbf{w}$ との内積を最大となるのは，求める射影軸が $\sum_{k=1}^{N}(N-2k+1)\mathbf{z}_{r(k)}$ と同じ方向を向くときである．従って，

$$\hat{\mathbf{w}} \propto \sum_{k=1}^{N}(N-2k+1)\mathbf{z}_{r(k)} \tag{3}$$

を得る．

2.2 アルゴリズム

目的関数を最大化するには，射影後の値により降順ソートする必要があり，また，降順ソートするには，射影軸 $\mathbf{w}$ を求める必要があるため，反復法により目的関数を最適化する．目的関数 $\mathcal{R}$ を最大化するアルゴリズムを説明する．

入力：元空間の座標ベクトル群 $\mathbf{Z} = (\mathbf{z}_1, \cdots \mathbf{z}_N)$
(N は全オブジェクト数を示す)
：埋め込み次元 L

1. 初期化：s = 0, $\mathbf{w}_s$ をランダムに初期化；
2. 射影：$||\mathbf{w}_s|| = 1$ と正規化し，$\mathbf{x} = \mathbf{w}_s^T\mathbf{Z}$ で元空間座標ベクトル群 $\mathbf{Z}$ を $\mathbf{w}_s$ へ射影する；
3. ソート：射影後の値 $\mathbf{x}$ によりオブジェクトを降順ソートし，ソート後のオブジェクト順序リスト r を構築する；
4. 更新：射影軸を $\hat{\mathbf{w}}_{s+1} = \sum_{k=1}^{N}(N-2k+1)\mathbf{z}_{r(k)}$ と更新する．
5. 判定：目的関数値 $v = \hat{\mathbf{w}}_{s+1}^T\sum_{k=1}^{N}(N-2k+1)\mathbf{z}_{r(k)}$ が収束すれば $\mathbf{x} = \mathbf{w}^T\mathbf{Z}$ を出力し終了，さもなければ $s \leftarrow s+1$ とし 2. へ；
出力：埋め込み座標ベクトル群 $\mathbf{X} = (\mathbf{x}_1^T, \cdots \mathbf{x}_L^T)^T$

上述したアルゴリズムにより，射影後のバラつきが最大となるような射影軸 $\mathbf{w}$ を求めることができる．第 2 軸以降を求める際には，既に求めた射影軸 $\mathbf{w}$ の直交補空間において，アルゴリズムを実行することにより求めることができる．

2.3 頑健線形射影法の特徴

RLP 法には以下の 3 点の特徴があげられる．まず始めに絶対値を距離として採用しているため，自乗値を距離としている PCA 法と比較して外れ値による影響が小さくなると考えられる．次に，ソートすることによりアルゴリズム全体の計算量を軽減している点である．単純に全オブジェクト間の距離（類似度）を計算するには，オブジェクト数の自乗オーダー（$O(N^2)$）の時間計算量がかかる．提案アルゴリズムでは，射影後の値でオブジェクトを降順ソートし，オブジェクト数のオーダーで目的関数値を計算できるため，アルゴリズム全体の時間計算量はソートにかかる $O(N\log N)$ となり，効率的なアルゴリズムになっている．最後に RLP 法は，反復ごとに射影軸 $\mathbf{w}$ の改善が数学的に保証されているため，射影軸 $\mathbf{w}$ について収束の保証性が期待できる．

3. 実験による評価

RLP 法，PCA 法, ランダム射影法に対して，以下で説明するレビューサイトのアイテム集合を用いて評価する．

3.1 実験データ

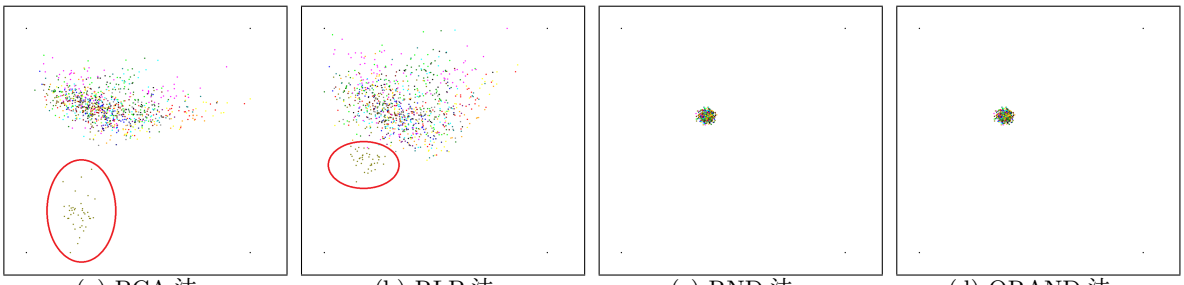
本稿では@cosme，アニコレを実験データとして用いた．@cosme [5] は，化粧品をアイテムとしたレビューサイトであり，2008 年 12 月から 2009 年 12 月にかけてクロールして取得したものである．48,548 アイテム，45,024 ユーザー，331,084 レビューを有する．本稿ではレビュー数の多い上位 1,000 アイテムを評価対象とした．

アニコレ [6] は，アニメをアイテムとしたレビューサイトであり，2012 年 8 月 8 日に取得したものである．1,790 アイテム，13,111 ユーザー，299,887 レビューを有する．

アイテム数を N，ユーザ数を M とし，アイテム i へのユーザー評点を要素とする M 次元ベクトル $\mathbf{z}_i$ で定義し，$||\mathbf{z}_i|| = 1$ と正規化して入力とした．

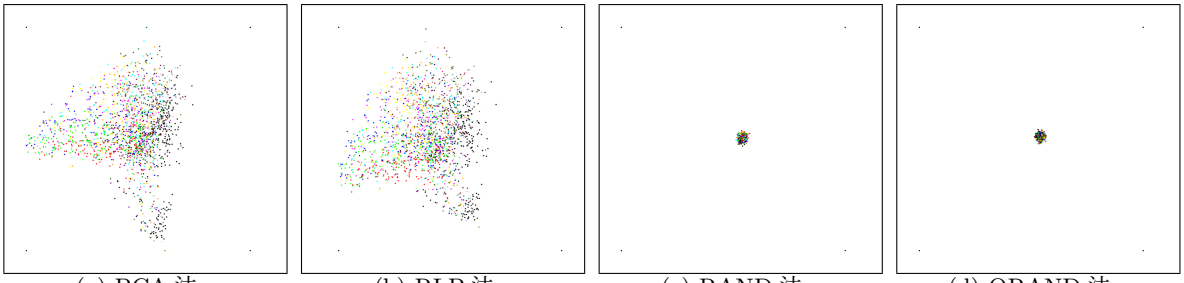
3.2 ランダム射影法

ランダム法では求める射影軸をランダムに設定し，ランダム設定された射影軸に入力データである元空間座標ベクトル群を射影する．なお，ランダム法では PCA 法,RLP 法とは異なり目的関数値の最大化は行わない．本稿ではランダム設定され



(a) PCA 法　(b) RLP 法　(c) RND 法　(d) ORAND 法

図 1: 2 次元平面布置結果（@cosme）



(a) PCA 法　(b) RLP 法　(c) RAND 法　(d) ORAND 法

図 2: 2 次元平面布置結果（アニコレ）

る射影軸同士が直交する場合と, 直行しない場合の 2 つのケースを用意し，直交化付きランダム法を ORAND 法，直交化なしランダム法を RAND 法とする．

3.3　目的関数値による評価

本研究において，目的関数値は全オブジェクト間の距離の総和であり，その次元空間におけるデータ量と呼ぶ．次元圧縮後の目的関数値が大きければ大きいほど元空間である高次元空間でのデータ量に近づくため，目的関数値が大きい方がより高次元空間でのデータ量を保存したといえる．表の目的関数値 $L1$ は距離計算にマンハッタン距離を用いた結果を表し $L2$ は距離計算にユークリッド距離を用いた結果を表す．

表 1，表 2，表 3，表 4 を見ると，両実験データでも PCA 法は距離計算にユークリッド距離を用いているため $L2$ においてはどの手法よりも目的関数値が大きい結果となった．一方,RLP 法は距離計算にマンハッタン距離を用いているため $L1$ においては PCA 法を上回り, どの手法よりも目的関数値の大きい結果となった．しかしながらランダム法は $L1$，$L2$ 両者とも他の方法と比較すると非常に小さな結果となった．そのためランダム法は他の方法に比べて高次元空間でのデータ量を保存していない縮小写像結果となったといえる．

3.4　可視化結果による評価

@cosme の分析結果 図 1(a) と 図 1(b) を比較してみると，@cosme ではどちらの手法でもスキンケア用品，メイクアップアイテムなど使用用途ごと，また同一のブランドごとにまとまって近傍に射影されている．このデータでは”ラッシュ”というブランドのアイテムがほかのブランドとは異なるユーザー層に支持されており他のブランドとの距離が大きくなっている．そのため，自乗距離を最大化する PCA 法では”ラッシュ”とその他のブランド間の距離が支配的となり他のブランド間の距離が縮小された結果となったと考えられる．RLP 法ではこのような点に対応するため絶対値距離を採用しているため全体が万遍なく散らばり，より解釈可能な可視結果が得られた．一方ランダム法は分析結果 図 1(c) と 図 1(d) を見てみると両者ともデータが散らばることなく密集した結果となった．これは目的関数値が PCA 法,RLP 法に比べて非常に小さく，そのためデータが散らばり具合が PCA 法，RLP 法に比べて非常に小さくなったのだと考えられる．

アニコレの分析結果 図 2(a) と 図 2(b) を比較してみると，どちらの手法でも，子供向け映画に関するアイテムや原作に関するアイテム同士などが近傍に射影されており，もとの高次元空間での類似構造を保存した射影結果が得られた．こちらもランダム法による結果 2(c) と 図 2(d) を見てみると両者ともデータが密集した結果となった．

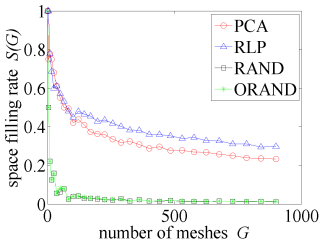
3.5　空間充填率による評価

空間充填率を以下のように定義し，2 次元可視化結果を定量的に評価する．2 次元平面を G 個のメッシュに分割する．g 番目のメッシュにオブジェクトが含まれていれば $c_g = 1$ とし，空間充填率を以下のように計算する．

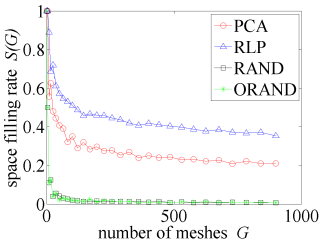
$$S(G) = \frac{1}{G}\sum_{g} c_g \tag{4}$$

明らかに，オブジェクトが 2 次元平面上にあまねく散らばっている方が $S(G)$ が高くなる．横軸がメッシュの数 G，縦軸が空間充填率 $S(G)$ とする．

@cosme の分析結果，図 3(a) を見ると，ランダム法は両者とも目的関数値が小さいため，データが密集しメッシュの数に関わらず非常に空間充填率の低い結果となった．一方,RLP 法はメッシュを増加していくと PCA 法に比べて常に約 20 %ほど高い空間充填率を示し，RLP 法の方が PCA 法より射影後の散らばり具合が大きく空間充填率の高い射影が実現できている．

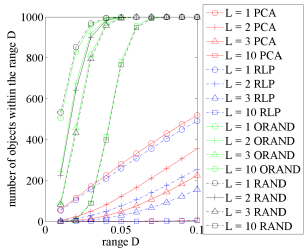


(a) 空間充填率（@cosme）

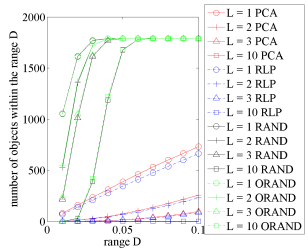


(b) 空間充填率（アニコレ）

図 3: 空間充填率結果



(a) レンジクエリ結果（@cosme）



(b) レンジクエリ結果（アニコレ）

図 4: レンジクエリ結果

アニコレの分析結果，図 3(b) を見るとランダム法は両者とも空間充填率が低く，データの散らばり具合が小さい結果となり，RLP 法はメッシュを増加していくと PCA 法に比べて常に約 10 %ほど高い空間充填率を示した．こちらも RLP 法の方が空間充填率の高い，よりデータの散らばった可視化結果となった．

3.6 類似検索機能による評価

低次元空間への射影結果に関して，類似検索の観点から定量的に評価する．一般に，低次元空間の方が高次元空間での距離計算よりコストが少なく済む．縮小写像により射影した低次元空間で範囲 D 内にあるオブジェクトは，高次元空間でも範囲 D 内に存在する．従って，元空間（高次元空間）で範囲 D 内に存在するオブジェクトを検索するレンジクエリにおいて，低次元空間で範囲 D 内にあることは必要条件であるため，条件に満たないオブジェクトを枝刈りできれば理想的である．

オブジェクト i から範囲 D 内に存在するオブジェクトの集合を $Q_D(i)$ としたとき，全オブジェクトの平均 D 近傍オブジェクト数 $\bar{Q}_D = \frac{1}{N}\sum_{i=1}^{N}|Q_D(i)|$ により評価する．RLP 法ならびに PCA 法, ランダム法は縮小写像であるため，元空間における範囲 D 内の平均オブジェクト数より，低次元射影後における範囲 D 内の平均オブジェクト数の方が多くなる．しかしながら，低次元射影後にできるだけ少ない数のオブジェクトに絞ることができれば，レンジクエリ検索の効率化が期待できる．

図 4(a) と図 4(b) に類似検索の観点から，射影後の低次元空間での平均近傍オブジェクト数による定量評価結果を示す．検索用データベース内に，既に非常に類似するオブジェクトが存在するかを確認するというタスクの観点から，0.01 から 0.1 の範囲でレンジを変化させた結果を示す．

横軸をレンジ D，縦軸を平均近傍オブジェクト数 $\bar{Q}_D$ とする．@cosme の分析結果である図 4(a) を見ると，10 次元以上への射影では有意な差は見られないものの，低次元へ射影した空間において，PCA 法より RLP 法の方が平均近傍オブジェクト数が小さいことがわかる．$L = 1$，つまり 1 次元では大きな差は見られなかったが，$1 < L \leq 3$ 次元では有意な差が確認できた．すなわち，RLP 法を用いた射影による低次元での距離計算において，より少ないオブジェクト数に絞ることができ，効率的なレンジクエリ検索への貢献が期待できる．一方，ランダム法は両者ともレンジの増加に伴い, 平均近傍オブジェクト数が急激に増加する結果となった．これは目的関数値が小さいがためにデータの散らばり具合が小さくなり，そのためレンジが増加すると一気に平均近傍オブジェクト数が増加したと考えられる．

アニコレの分析結果，図 4(b) を見ると，@cosme ほど顕著な差ではないが，RLP 法による射影空間での平均近傍オブジェクト数の方がどの手法よりも小さいことがわかる．こちらも距離に絶対値を採用したために近傍に射影されていたアイテム同士が散らばり，範囲 D 内の平均オブジェクト数が減少したと考えられる．

4. おわりに

本論文では，外れ値に対して頑健性を示し，元空間でのオブジェクト間の関係をできるだけ保存する埋め込み法について研究し，射影後のデータの散らばり具合が PCA 法，ランダム法よりも大きくなる埋め込み法を提案した．また，データによっては，外れ値の影響を受けて効果的な枝刈りが期待できない可能性があったが，提案法により，データの散らばり具合が向上したので，類似検索において非類似オブジェクトの効果的な枝刈りが期待できることを確認できた．今回比較に用いたランダム法は目的関数値の向上を行わないために目的関数値が小さい結果となった．そのため予想と反して，データが散らばらず RLP 法，PCA 法に比べてデータの密集した結果となってしまった．今後はピボット法を用いた類似検索機能の面から提案法を評価し，その有効性を検証していく．

謝辞 本研究は，科学研究費補助金基盤研究 (C)(No.23500128) の補助を受けた．

参考文献

[1] J. A. Lee, and M. Verleysen.: ”Nonlinear Dimensionality Reduction (Information Science and Statistics)” Springer-Verlag, New York, (2008)

[2] R. Spence, A. Press.: ”Information Visualization”, Addison-Wesley, 2001.

[3] P. Zezula, G. Amato, V. Dohnal, and M. Batko, “Similarity Search: The Metric Space Approach (Advances in Database Systems)”, Springer-Verlag, New York, 2006.

[4] B. Bustos, G. Navarro, and E. Chavez.: ”Pivot Selection Techniques for Proximity Searching in Metric Spaces”, Proc. of Pattern Recognition Lettes, Vol.24, No.14, pp. 2357-2366, 2003.

[5] http://www.cosme.net/

[6] http://www.anikore.jp/